THE FALLIBILITY PARADOX*

By Chandra Sripada

Abstract: *Reasons-responsiveness theories of moral responsibility are currently among the most popular. Here, I present the fallibility paradox, a novel challenge to these views. The paradox involves an agent who is performing a somewhat demanding psychological task across an extended sequence of trials and who is deeply committed to doing her very best at this task. Her action-issuing psychological processes are outstandingly reliable, so she meets the criterion of being reasons-responsive on every single trial. But she is human after all, so it is inevitable that she will make rare errors. The reasons-responsiveness view, it is claimed, is forced to reach a highly counterintuitive conclusion: she is morally responsible for these rare errors, even though making rare errors is something she is powerless to prevent. I review various replies that a reasons-responsiveness theorist might offer, arguing that none of these replies adequately addresses the challenge.*

KEY WORDS: moral responsibility, reasons-responsiveness views, lottery paradox, valuationist views, slips, errors

I. Introduction

A number of philosophers say that moral responsibility requires satisfying a reasons-responsiveness condition.[1] One influential version of this view says the condition is satisfied when, very roughly, in a suitably broad range of scenarios in which there is sufficient reason to do otherwise, the mechanism that issues in action issues in alternative actions.[2] In this essay, I discuss a problem for these views.

* Thanks to the contributors to this volume for extensive feedback on an earlier draft of this essay. Special thanks to Michael McKenna, Samuel Murray, Manuel Vargas, and an anonymous reviewer for this journal for detailed comments that greatly improved the manuscript.

[1] Reasons-responsiveness is typically offered as a necessary, but not sufficient, condition for moral responsibility. Other common criteria include a knowledge condition and historical condition, among others. I am assuming throughout this essay, unless noted otherwise, that these other conditions for moral responsibility are met.

[2] Another common formulation is agent-based rather than mechanism-based: an agent is morally responsible for an action only if the agent is reasons-responsive. See David O. Brink and Dana K. Nelkin, *Fairness and the Architecture of Responsibility*, ed. David Shoemaker (Oxford: Oxford University Press, 2013); Manuel Vargas, *Building Better Beings: A Theory of Moral Responsibility* (Oxford: Oxford University Press, 2013); Michael McKenna, "Reasons-Responsiveness, Agents, and Mechanisms," in *Oxford Studies in Agency and Responsibility Volume 1*, ed. David Shoemaker (Oxford: Oxford University Press, 2013), 151–83. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199694853.001.0001/acprof-9780199694853-chapter-7, for more on the distinction). For ease of exposition, I formulate the fallibility paradox for mechanism-based reasons-responsiveness views first. Later, in Section V.C, I argue that switching to an agent-based formulation makes little difference.

doi:10.1017/S0265052519000293

II. A Puzzle

Fei, an undergraduate student, signs up for a psychological experiment. The researchers give her a version of the Stroop task, a classic task used in countless studies.[3] On each trial, a color word is presented on a screen (for example, "red," "blue," "green") and the word itself is presented in an ink color. Fei is instructed that on all the trials, she should respond with the ink color of the word and not the word meaning (she responds by pressing one of five buttons, each associated with a color). On some trials, the color word presented on the screen matches its ink color, making it easier to get the correct response. On other trials, they don't match (for example, the word "red" is shown in yellow colored ink), and it is more challenging to produce the correct response.

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 errors. By "error," what I mean is that she intentionally presses one of the five buttons for her response but the button she presses is incorrect.[4] Overall, her response accuracy is very impressive: she is in the top 0.01% of people who have taken this task.[5]

[3] See C. M. MacLeod, "Half a Century of Research on the Stroop Effect: An Integrative Review," *Psychological Bulletin* 109, no. 2 (1991): 163–203, for a review of the history of this task and summary of key findings.

[4] In a series of papers, Santiago Amaya has drawn attention to *slips*, which are quite similar to what I am here calling "errors." On his account, slips are to be understood as intentional actions that fail to correspond to what the agent preferred to do at the time. He also distinguishes slips from other kinds of agential failings, such irresoluteness and "Freudian" conduct. See Amaya Santiago, "Slips," *Noûs* 47, no. 3 (2013): 559–76. https://doi.org/10.1111/j.1468-0068.2011.00838.x; Amaya, "The Argument from Slips," in *Agency, Freedom, and Moral Responsibility*, ed. Andrei Buckareff, Carlos Moya, and Sergi Rosell (London: London: Palgrave Macmillan, 2015), 13–29, as well as Santiago Amaya and John M. Doris, "No Excuses: Performance Mistakes in Morality," in *Handbook of Neuroethics*, ed. Jens Clausen and Neil Levy (Dordrecht, Netherlands: Springer, 2015), 253–72.

[5] The standard finding is that subjects make an error on roughly 5-10 percent of the "incongruent" trials where the named color and ink color disagree (MacLeod, "Half a Century of Research on the Stroop Effect"), and error rates go down if incentives are given for accurate responding (Mimi Liljeholm and John P. O'Doherty, "Anything You Can Do, You Can Do Better: Neural Substrates of Incentive-Based Performance Enhancement." PLoS Biology 10, no. 2 (2012): e1001272. https://doi.org/10.1371/journal.pbio.1001272.) However, across the course of a prolonged experiment, virtually no one, no matter what incentives are given, achieves perfect accuracy (indeed, as incentives get sufficiently high, performance often suffers due to "choking under pressure" effects; Dan Ariely, Uri Gneezy, George Loewenstein, and Nina Mazar, "Large Stakes and Big Mistakes," *The Review of Economic Studies* 76, no. 2 (2009): 451–69. https://doi.org/10.1111/j.1467-937X.2009.00534.x; Vikram S. Chib, Shinsuke Shimojo, and John P. O'Doherty, "The Effects of Incentive Framing on Performance Decrements for Large Monetary Outcomes: Behavioral and Neural Mechanisms," *Journal of Neuroscience* 34, no. 45 (2014): 14833–44. https://doi.org/10.1523/JNEUROSCI.1491-14.2014.

Now, consider this question: On each trial, does Fei's action arise from a reasons-responsive mechanism? The answer seems to be: surely yes. On each trial, Fei performs an intentional action: reaching out for and pressing one of the buttons. A number of psychological processes contribute to her performing this action. As we shall see later in more detail, these include processes related to attention, decision-making, and intention formation. The mechanism that issues in Fei's action is naturally understood in terms of the joint operation of these processes; this mechanism helps her to both track what the correct response is and produce the correct response.

Of course, this mechanism does not operate flawlessly, since she does make some errors. But reasons-responsiveness theorists almost never insist that to qualify as reasons-responsive, the relevant mechanism that issues in action must be unerring in tracking one's reasons. Rather, it is typical to set some threshold for reasons tracking, often a fairly lenient one. For example, theorists typically say the relevant mechanism must issue in alternative actions in "at least one" or "some" of the worlds in which there is sufficient reason to do otherwise.[6]

In Fei's case, the reasons that she has to produce one of the five responses change constantly throughout the task as she is shown different color words printed in different ink colors. The mechanism that issues in Fei's actions in the task is extraordinarily adept at tracking these reasons, which is why she gets the correct response on more than 99.5% of the trials, one of the best performances on record. Thus, it certainly seems plausible that in her case, the mechanism that issues in action in this task meets the threshold for being reasons-responsive.[7]

Now, if Fei acts on a reasons-responsive mechanism on each trial of the task, and if we assume a reasons-responsiveness view of moral responsibility is correct (and there are no defeaters pertaining to the other criteria for moral responsibility), then it follows that on each trial, Fei is morally responsible for what she does on that trial. That is, she is morally responsible for what she does on trial 1, morally responsible for what she does on trial 2, morally responsible for what she does on trial 3, and so on for all 1,000 trials.

[6] Fischer says the mechanism that issues in action must be "moderately" reasons responsive: across worlds in which there is sufficient reason to do otherwise, it "regularly" recognizes these reasons and reacts to these reasons in at least one world (John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* [New York: Cambridge University Press, 1998]). Others set a higher threshold for reasons reactivity (e.g., Brink and Nelkin, *Fairness and the Architecture of Responsibility*.) See McKenna, "Reasons-Responsiveness, Agents, and Mechanisms" for more general discussions of the role of thresholds in reasons-responsiveness accounts.

[7] Thus far, I am assuming that, since all 1,000 trials are highly similar and she tries equally hard on every trial, it is the same mechanism that issues in action across all the trials. This strikes me as a highly intuitive picture of what goes on in this task. In Section IV, I consider the possibility that different mechanisms are at work on correct versus incorrect trials.

But now, it seems, we have a problem. Given all that we know from thousands of studies of performance on the Stroop (and Stroop-like) tasks, just about *everyone* will make at least a few errors on this task, and this isn't something that they can avoid. Fei, in particular, due to her deep love for animals, tried very hard to get the correct answer on *all* the trials, and she performed extraordinarily well: she is in the top 0.1% of those who have taken the task. But even she did not perform flawlessly.

It appears to be a basic feature of human psychology, just the way that our minds are set up, that errors of the kind that Fei made occasionally occur. We can try harder to focus on a task and thus drive down the rate of errors somewhat. But for us humans, such errors cannot be eliminated (and this is for principled reasons, as I discuss in what follows).[8] If the occurrence of rare errors on Stroop-type tasks is not something that we humans have the power to prevent, then, for a person who doesn't want such errors to happen and sincerely tries to prevent them, it seems plainly wrong to say she is nonetheless morally responsible for them.

Yet, because Fei's actions arise from a reasons-responsive mechanism throughout the task, the reasons-responsiveness view says that Fei *is* morally responsible for what she does in *all* the trials including her rare errors. The chain of reasoning that leads to this inconsistency is what I call the "fallibility paradox."[9]

[8] Elsewhere, I discuss the connection between inevitable rare errors of the kind Fei exhibits and loss of control in non-laboratory, "real-world" contexts; see Chandra Sripada, "Addiction and Fallibility" *Journal of Philosophy* 115, no. 2 (2018): 569–87; Sripada, "Self-Expression: A Deep Self Theory of Moral Responsibility," *Philosophical Studies* 173, no. 5 (2016): 1203–32. https://doi.org/10.1007/s11098-015-0527-9.

[9] Slips and errors have figured into other challenges to reasons-responsiveness views. But interestingly, the challenge posed by the fallibility paradox comes from a diametrically opposed direction. The fallibility paradox presents a challenge for reasons-responsiveness views because it says these views are *overinclusive*: they count an agent as morally responsible for certain kinds of slips and errors when they should not. Reasons-responsiveness views have also faced the opposite charge: it is claimed they are *underinclusive* and fail to count an agent as morally responsible for certain kinds of slips and errors when they should. Reasons-responsiveness theorists have offered responses. For example, McKenna and Warmke discuss the findings from the literature on situationism. They consider the claim that these findings show we are typically not reasons-responsive enough, opening the door to skepticism about moral responsibility, and they offer detailed replies. Samuel Murray, "Responsibility and Vigilance," *Philosophical Studies* 174, no. 2 (2017): 507–27. https://doi.org/10.1007/s11098-016-0694-3, considers cases of forgetting and other failures of vigilance. He puts forward an argument for why a reasons-responsiveness view can in fact account for why people are morally responsible for these failings. However, even if McKenna and Warmke (Michael McKenna and Brandon Warmke, "Does Situationism Threaten Free Will and Moral Responsibility?" *Journal of Moral Philosophy* 1, no. 36 [2017]) and Murray are right and the charge of underinclusiveness is successfully rebutted, this does not address the fallibility paradox, which attacks from precisely the opposite direction.

III. Relation to the Lottery Paradox

Some readers will have noticed that the fallibility paradox bears some similarities to the lottery paradox, first set out by Henry Kyburg,[10] one version of which goes as follows: Let us suppose that it is rational to accept a proposition if its probability of being true is very high, say at least 99.5% likely. Now consider a fair lottery with 1,000 tickets and a single winner. By our criteria, it is clearly rational to accept that the first ticket will not win. The same is true of the second ticket. And the third ticket. And so on for all 1,000 tickets. Given these observations, it seems we are entitled to conclude something stronger, based on the following inference:

> (**A**) If it is rational to accept that for each ticket, that ticket will not win, then it is rational to accept that no ticket will win.

This, of course, raises a problem. You already know that the fair lottery has a single winner, and so it is clearly rational to accept that one ticket will win. How can it be rational to accept both that a single ticket will win and that no ticket will win?

The core similarity between the lottery paradox and the fallibility paradox is that they challenge the use of seemingly arbitrary *thresholds* in philosophical accounts of rational acceptance and moral responsibility, respectively. They show that no matter how the thresholds are set, counterexamples emerge.

In the case of the lottery paradox, the threshold concerns rational acceptance. Some theorists might insist on a stringent threshold, say 0.995 probability, before one should rationally accept a proposition. Others might propose a more lenient standard. In the case of reasons-responsiveness views of moral responsibility, the threshold concerns the quantity of possible worlds in which the action-issuing mechanism does otherwise where there is sufficient reason to do so. As noted earlier, Fischer says the mechanism must do otherwise in at least one of these possible worlds, while other theorists have proposed more stringent thresholds.

Regardless of where the thresholds are set, for both accounts, once the thresholds are crossed, that is good enough: the proposition counts as rationally accepted and the mechanism counts as reasons-responsive, respectively.

Given the role of thresholds in these accounts, the two paradoxes produce their puzzling conclusions by applying the respective concepts (that is, rational acceptance and moral responsibility) in "iterated contexts." These are contexts that involve a large number of highly similar cases. We antecedently know that the relevant concept does not apply to all the

[10] Henry Kyburg, *Probability and the Logic of Rational Belief* (Middletown, CT: Wesleyan University Press, 1961).

cases, yet in each individual case, the relevant concept's associated threshold is exceeded. So the preceding accounts tell us to apply the relevant concept to all the cases. This results in the puzzling inconsistencies that lie at the heart of both paradoxes.

There are also differences between the fallibility paradox and the lottery paradox. Let me highlight one particularly important one. The chain of reasoning in the lottery paradox relies on a principle of *agglomeration*, one version of which says: for propositions $p_1, p_2, \ldots p_n$, if it is rational to accept that p_1, and if it is rational to accept that p_2, and so on up to p_n, then it is rational to accept (p_1 & p_2, . . . p_n). This principle underlies the inference laid out in **(A)** above, which is an essential step in setting up the lottery paradox.

Agglomeration principles are problematic, and Kyberg's strategy for addressing the lottery paradox involved disallowing the agglomerative step.[11] The fallibility paradox, however, does not rely on agglomeration at all. You can see this by looking closely at this step of the argument:

> If Fei acts on a reasons-responsive mechanism on each trial of the task, and if we assume a reasons-responsiveness view of moral responsibility is correct (and there are no defeaters pertaining to the other criteria for moral responsibility), then it follows that on each trial, Fei is morally responsible for what she does on that trial. That is, she is morally responsible for what she does on trial 1, morally responsible for what she does on trial 2, morally responsible for what she does on trial 3, and so on for all 1,000 trials.

This premise is not agglomerative. An agglomerative version of this premise would say that if Fei is morally responsible for what she does on trail 1, morally responsible for what she does on trial 2, morally responsible for what she does on trial 3, and so on, then she is morally responsible for (what she does on trial 1 & what she does on trial 2 & what she does on trial 3, and so on . . .). The thing in the parentheses is what we might call a "compound action"; it is the conjunction of multiple actions, and it is analogous to the conjunction of beliefs that plays a key role in the lottery paradox.

Importantly, the fallibility paradox does *not* rely on attributing moral responsibility for compound actions. Indeed, it is not even clear if such attributions are sensible. Leaving this question aside, the point I want to emphasize is that the fallibility paradox does not rely on agglomeration in any form. This observation significantly strengthens the fallibility paradox as a counterexample to reasons-responsiveness views; one cannot challenge the paradox by pointing to the problematic nature of agglomeration.

[11] Ibid.

IV. The Different Mechanisms Response

Consider the following reply from a reasons responsiveness theorist: The mechanism that issues in action in Fei's 996 correct trials is a reasons-responsive mechanism and she is morally responsible for what she does on those trials. But on the four trials in which she produces an incorrect response, a different type of mechanism operates. Furthermore, *this* mechanism is not reasons-responsive. In this way the reasons-responsiveness theorist can capture the intuition that she is not, after all, morally responsible for what she does on these four trials.

To make some progress in evaluating the "different mechanism" proposal, we need a detailed specification of what are the psychological mechanisms that operate in the Stroop task.[12] Here, we are in luck. The Stroop task is one of the most extensively studied in psychology and neuroscience. Based on this body of work, I want to now present a model of Stroop task performance with an eye toward illuminating whether the "different mechanism" response is viable.[13]

A. The psychology and neuroscience of Stroop task performance

According to a highly influential model, the Stroop task situation involves a competition between two candidate responses: a relatively habitual "default" response (word reading) and an alternative response (color naming) that requires top-down support to be implemented. These two responses compete to win selection by decisional systems, which in turn produce one's proximal intentions to act.

Word reading is the response favored by fast-operating habitual systems. These systems are specialized for using certain forms of iterative reward learning to favor situationally adaptive responses.[14] For nearly all of us in modern societies, reading words is extensively practiced over the course

[12] Fischer says very little about how to individuate mechanisms for the purposes of evaluating their reasons-responsiveness, and in nearly all his examples, the relevant mechanism is not specified in any detail. McKenna ("Reasons-Responsiveness, Agents, and Mechanisms") takes up this problem for Fischer's view in some detail.

[13] There are a number of computational models of Stroop task performance, all of which are compatible with the key conclusions I want to draw (e.g., R. Hans Phaf, H. C. Van der Heijden, and Patrick T. W. Hudson,"SLAM: A Connectionist Model for Attention in Visual Selection Tasks," *Cognitive Psychology* 22, no. 3 [1990]: 273–341. https://doi.org/10.1016/0010-0285(90)90006-P; Ardi Roelofs, "Goal-Referenced Selection of Verbal Action: Modeling Attentional Control in the Stroop Task," *Psychological Review* 110, no. 1 [2003]: 88–125.) I am here focusing on the approach of Jonathon Cohen and his colleagues as laid out in a number of articles (see, for example, J. D. Cohen, K. Dunbar, and J. L. McClelland, "On the Control of Automatic Processes: A Parallel Distributed Processing Account of the Stroop Effect," *Psychological Review* 97, no. 3 [1990]: 332–61.)

[14] See Ray J. Dolan and Peter Dayan, "Goals and Habits in the Brain," *Neuron* 80, no. 2 (2013): 312–25. https://doi.org/10.1016/j.neuron.2013.09.007, for a review of brain-based computational algorithms that underlie habit learning.

of years, and based on this learning history, habitual systems strongly favor this candidate response.

Color naming, in contrast, is not highly practiced, and thus it is initially in a weaker position relative to word reading. However, based on the explicit instructions given to subjects that color naming is the correct response, subjects can use *regulatory processes* to bias the competition between the habitually favored response and the alternative.

A standard model[15] of how regulation works in the Stroop task is through deployment of top-down attention to enhance perceptual representations of the stimulus dimension associated with correct responding (that is, the color of the stimulus) and dampen perceptual representations associated with irrelevant stimulus dimensions (geometric shape of the letters, which is the basis of reading). Deployment of top-down attention in this way strongly raises the probability that subsequent decision and intention formation processes will produce the correct response.

Notice the effect of top-down attention is *probabilistic*. This is a central feature of most models of Stroop task performance. Top-down attention directed to the task-relevant stimulus dimensions strongly raises the probability that the task-appropriate response will be selected and executed but does not guarantee it. Why is the effect only probabilistic?

The standard answer is that human information processing involves an ineliminable role for stochasticity; this is an inherent feature of the representational formats and primitive operations utilized by the brain.[16] On each trial, during the dynamic evolution of the competition between the habitual response and the top-down supported response, countless stochastic influences are continuously buffeting the underlying neural representations/operations that realize the competition between the responses. These "noise" processes subtly and continuously alter the

[15] See, for example, Tobias Egner and Joy Hirsch, "Cognitive Control Mechanisms Resolve Conflict through Cortical Amplification of Task-Relevant Information." *Nature Neuroscience* 8, no. 12 (2005): n.1594. https://doi.org/10.1038/nn1594.

[16] For discussions of stochasticity in neural computation, see Michael N. Shadlen and Adina L. Roskies, "The Neurobiology of Decision-Making and Responsibility: Reconciling Mechanism and Mindedness," *Frontiers in Neuroscience* 6 (2012). https://doi.org/10.3389/fnins.2012.00056; Michael N. Shadlen, "Comments on Adina Roskies,'Can Neuroscience Resolve Issues about Free Will?'" In *Moral Psychology, Volume 4: Free Will and Moral Responsibility*, ed.Walter Sinnott-Armstrong (Cambridge, MA: MIT Press, 2014), 39–50. For discussions of how stochasticity manifests in performance in Stroop-like tasks, see Arthur R. Jensen, "The Importance of Intraindividual Variation in Reaction Time," *Personality and Individual Differences* 13, no. 8 (1992): 869–81. https://doi.org/10.1016/0191-8869(92)90004-9; John R. Nesselroade and Nilam Ram, "Studying Intraindividual Variability: What We Have Learned That Will Help Us Understand Lives in Context," *Research in Human Development* 1, nos. 1-2 (2004): 9–29. https://doi.org/10.1080/15427609.2004.9683328; F. Xavier Castellanos, Edmund J. S. Sonuga-Barke, Anouk Scheres, Adriana Di Martino, Christopher Hyde, and Judith R. Walters,"Varieties of Attention-Deficit/Hyperactivity Disorder-Related Intra-Individual Variability," *Biological Psychiatry* 57, no. 11 (2005): 1416–23. https://doi.org/10.1016/j.biopsych.2004.12.005.

strengths of the respective responses. As a consequence, the results of the competition are slightly unpredictable. With sufficient top-down attention, most of the person's responses will be correct, but a few will still be, due to stochasticity in the system, errors.[17]

Going forward, I refer to errors of the type Fei exhibited that are due to stochasticity alone as "noise-based" errors. It bears emphasis that I am not saying that noise-based errors are the only kind of errors—that is surely false. For example, sometimes people are sloppy or don't care much about a slip or two (or three), and these factors causally contribute to the occurrence of errors. Such errors aren't noise-based in the sense that I have in mind. Rather, my point is that (purely) noise-based errors *do* exist, and they offer a particularly potent way to formulate the fallibility paradox as a challenge to reasons-responsiveness views.

B. Evaluating the "different mechanisms" response

With the preceding detailed picture of Stroop task performance in mind, let us return to the "different mechanisms" response. In my view, this response is now seen to be deeply implausible. On every single trial of Fei's Stroop task, there seems to be a single mechanism at work. That overall mechanism consists in: 1) the processes that underwrite the habitual candidate response; 2) the processes that underwrite top-down attention and the candidate response it favors; 3) countless noise processes continuously influencing the unfolding competition between the two.

Importantly, in this overall mechanism, noise plays an ineliminable role in how response competition unfolds, and thus how the mechanism operates. As such, the possibility of error is *intrinsic* to the nature of the

[17] The main features of the preceding mechanistic description of Stroop task performance are nicely captured in the classic *drift diffusion model* of Roger Ratcliff and Gail McKoon, "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks," *Neural Computation* 20, no. 4 (2007): 873–922. https://doi.org/10.1162/neco.2008.12-06-420; Andreas Voss, Markus Nagler, and Veronika Lerche, "Diffusion Models in Experimental Psychology," *Experimental Psychology* 60, no. 6 (2013): 385–402. https://doi.org/10.1027/1618-3169/a000218). The model treats one's responses on a broad array of tasks as arising from a continuous random diffusion process (called Weiner-type diffusion) that evolves over time, eventually hitting a decision boundary that determines the response (READ the word or say the INK color). Top-down attention serves to strongly bias the evolution of the diffusion process in favor of the correct response (INK). But in each time instant, noise processes can potentially push the evolving diffusion path in either direction. The result is that—so long as the level of top-down attention is sufficient—on most trials, the person produces the correct response, but, inevitably, rare incorrect decisions and subsequent responses will also occur. Certain modifications of the classic drift diffusion model are required for conflict tasks like the Stroop task (Rolf Ulrich, Hannes Schröter, Hartmut Leuthold, and Teresa Birngruber, "Automatic and Controlled Stimulus Processing in Conflict Tasks: Superimposed Diffusion Processes and Delta Functions," *Cognitive Psychology* 78 [May 2015]: 148–74. https://doi.org/10.1016/j.cogpsych.2015.02.005; Corey N. White, Mathieu Servant, and Gordon D. Logan, "Testing the Validity of Conflict Drift-Diffusion Models for Use in Estimating Cognitive Processes: A Parameter-Recovery Study," *Psychonomic Bulletin and Review* 25, no. 1 [2018]: 286–301. https://doi.org/10.3758/s13423-017-1271-2), but they don't change the preceding basic picture.

mechanism itself. If this is right, there is no good basis to single out the four trials in which Fei produces noise-based incorrect responses and say on those trials, a different mechanism operated.

The preceding point—that noise is intrinsic to the mechanisms underwriting Stroop performance—relates to a broader point about mechanism individuation. We normally individuate mechanisms in a way that allows that they sometimes fail, including for reasons due to pure noise. If a machine makes widgets and on rare occasions makes a defective one (say on round 504 and 2907), we say *the* machine, that is, a single entity, failed on those two occasions. We don't say there are two machines at work: machine #1 operates flawlessly and produced nearly all the widgets, while machine #2 produces only defective widgets and was at work on rounds 504 and 2907. It seems, then, that we should say something similar about the psychological processes at work while Fei performs the Stroop task: these interacting processes constitute a single overall mechanism and this mechanism produces 996 successful responses as well as four noise-based errors.

V. Other Responses by Reasons-Responsiveness Defenders

A. The knowledge condition

Standard reasons-responsiveness views say that in addition to acting from a reasons-responsive mechanism, moral responsibility requires that the person meet a *knowledge* requirement. Might this requirement help us explain why Fei is morally responsible for what she does on most trials of the Stroop task, but not on the small subset of trials in which she makes noise-based errors?

I believe this avenue isn't very promising. Fei knows a *lot* on each trial of the Stroop task: She knows what the researchers' instructions are, she knows what the ink color of the stimulus is, and she knows what she is doing when she presses a button (for example, when she presses the button associated with yellow, she knows that the button is associated with yellow). Moreover, on nearly all trials, she brings this knowledge to bear in producing the correct response. On a few trials, of course, the knowledge that she has isn't brought to bear on what she does effectively: for example, though she knows the ink color of the word on the screen is, say, yellow and though she knows she is supposed to respond with the ink color of the word throughout the task, she nevertheless responds with the color named—the incorrect response. But making this sort of error doesn't mean she *lacks* the relevant knowledge.

If we deny this claim—that is, if we say failing to bring to bear one's knowledge and lacking knowledge are the same thing for the purposes of assessing moral responsibility—then trouble awaits. A great number of careless or sloppy people will then be inappropriately let off the hook. These observations suggest appealing to the knowledge condition for

moral responsibility will not do much to help the reasons-responsiveness theorist address the fallibility paradox.

B. Morally responsibility versus blameworthiness

A reasons-responsiveness theorist might claim that, contra what has been assumed so far, Fei *is* in fact morally responsible for making the four errors on the Stroop task. She is not, however, necessarily blameworthy for making the errors—responsibility and blameworthiness are two different things. For example, if I am coerced into handing over the gold in the bank's safe, I am morally responsible for what I do (after extensive deliberation, I *chose* to give up the gold rather face the adverse consequences). But I may not be blameworthy, so long as the threat was sufficiently severe. In the end, though, I don't think this response does much to address the fallibility paradox.

At the heart of the fallibility paradox is a deeply intuitive principle that connects powerlessness with excuse from moral responsibility: a person cannot be morally responsible for something that she doesn't want to happen and is powerless to prevent. Given this core principle, we refuse to accept that Fei is morally responsible for her errors: Any human, no matter how committed she is to animals, would make at least a few errors; this is something she cannot prevent. Furthermore, given that she is not morally responsible for these errors, it follows she is not blameworthy for these errors, since in order for a person to be blameworthy for what she does, she must first be morally responsible for what she does.

The present strategy by the reasons-responsiveness theorist—to distinguish moral responsibility from blameworthiness—*does* capture the intuition that Fei is not blameworthy for her errors, and this is not nothing. Nonetheless, it still leaves the core issue that drives the fallibility paradox untouched. The reasons-responsiveness view still appears to have to say that Fei is morally responsible for her errors, despite the fact that these errors are something she did not want to happen and was powerless to prevent. *That* remains a serious problem for the view.

C. Switching to agent-based versions of reasons-responsiveness views

Thus far, I have formulated the fallibility paradox as a problem for the mechanism-based approach to reasons-responsiveness. Some theorists have argued for an agent-based version of the theory, in which responsiveness is a property of the whole agent rather than any particular mechanism.

Several arguments are usually made on behalf of the agent-based version of the view (the arguments that follow are drawn from McKenna[18]).

[18] McKenna, "Reasons-Responsiveness, Agents, and Mechanisms."

First, the agent-based view is said to be more natural and intuitive. Second, it is argued that the agent-based view can in fact deal with counterfactual intervener scenarios, which is something that Fischer touted as a specific benefit of the mechanism-based view. Third, it has been argued that it is not possible to provide a principled individuation of mechanisms, especially when several mechanism jointly interact in complex ways to produce an action. An agent-based view avoids this sort of difficulty.

It is an interesting and important question which version of a reasons-responsiveness view is overall better. But the point I want to highlight is that none of the putative benefits of an agent-based view relative to a mechanism-based view help much in dealing with the fallibility paradox.

Earlier, I argued that the fallibility paradox applies to any version of a reasons-responsiveness view that formulates the condition of reasons-responsivity in terms of a threshold (Section III). If an agent-based reasons-responsiveness view could avoid setting a threshold for reasons-responsivity, then it could avoid the fallibility paradox. But it is not at all clear how thresholds can be avoided.

Take the set of processes Y that make up an agent's psychology. Mechanism-based views assess the responsivity to reasons of some subset of Y, in particular the subset that was causally operative in producing the relevant action. A mechanism-based view must next select some level of responsivity. It is implausible that moral responsibility requires perfect responsivity, that is, the mechanism issues an alternative action in *every* world in which there is sufficient reason to do so. So mechanism-based theorists typically select some more lenient threshold, for example, the mechanism issues an alternative action in at least one or some worlds in which there is sufficient reason to do so.

Agent-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. But having made this specification of the target of assessment, they still face the exact same problem of setting a threshold. It is deeply implausible that moral responsibility requires that the agent perform an alternative action in *every* world in which there is sufficient reason to do so. This is why agent-based theorists too usually propose some more lenient standard.[19]

If this is right, switching to an agent-based formulation of a reasons-responsiveness view makes no difference. It is the threshold aspect of reasons-responsiveness views specifically that drives the fallibility paradox, and agent-based views must set thresholds just the same as mechanism-based views.

[19] Brink and Nelkin, *Fairness and the Architecture of Responsibility*.

VI. What the Fallibility Paradox Tells Us about Moral Responsibility

One way to get a deeper handle on what drives the fallibility paradox is to consider a family of views of moral responsibility that don't appear to be susceptible to the paradox.

So-called valuationist views[20] are an important family of views that contrast in major ways with reasons-responsiveness views. These views say that for an agent to be morally responsible for an action, the action must flow from and express her evaluative point of view. Valuationist views are typically formulated in terms of two parts, with different theorists filling in these parts in different ways. First, there is an account of which subset among an agent's attitudes constitutes her evaluative point of view: values, cares, policies, and so on. Second, there is an account of what it means for an action to flow from or express an agent's evaluative point of view. I will here focus on valuationist views that analyze the expression relation in terms of a type *causal contribution:* an action expresses the attitudes that constitute the agent's evaluative point of view if these attitudes causally contribute (in the right way) to the production of the action.[21]

Valuationist views—in particular the causal contribution variety (I drop the qualifier going forward)—appear to make the right predictions about the case of Fei used to formulate the fallibility paradox. Recall that Fei cares very much about animals; animal welfare is something she deeply values. On the 996 trials in which she gets the correct response, it is natural to say that her responses flow from her values; it is in virtue of valuing animals that she directs top-down attention to strongly favor the correct response, which in turns strongly causally contributes to her producing the correct response. On the four trials in which she produces the incorrect

[20] I borrow this helpful terminology from Samuel Murray, Elise D. Murray, Gregory Stewart, Walter Sinnott-Armstrong, and Felipe De Brigard, "Responsibility for Forgetting," *Philosophical Studies* (2018), 1–25. https://doi.org/10.1007/s11098-018-1053-3, who cite John M. Doris, *Talking to Our Selves: Reflection, Ignorance, and Agency*, reprint edition (New York: Oxford University Press, 2015) and Chandra Sripada, "Self-Expression: A Deep Self Theory of Moral Responsibility," *Philosophical Studies* 173, no. 5 (2016): 1203–32. https://doi.org/10.1007/s11098-015-0527-9 as recent examples of valuationist views. Murray and colleagues' article focuses on moral responsibility for slips and errors in cases where the agent *should* be morally responsible. The fallibility paradox, as I noted earlier, presents the opposite kind of challenge: it concerns slips where the agent should *not* be morally responsible. A complete defense of a valuationist approach to responsibility for slips and errors should address Murray and colleagues' arguments, though I will not attempt such a defense here.

[21] Sripada, "Self-Expression: A Deep Self Theory of Moral Responsibility" argues that the specific kind of causal contribution that is relevant for moral responsibility is *motivational contribution*: an action expresses an element of one's evaluative point of view if that element motivationally supports performing the action. Older valuationist views understood the idea of expressing or flowing from one's evaluative point of view in explicit, conscious, and often highly rationalistic terms. For example, an action expresses an agent's evaluative point of view only if the agent consciously, reflectively endorses the action. Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy* 68, no. 1 (1971): 5–20. https://doi.org/10.2307/2024717.

response, her actions don't flow from her values. On each of these trials, she—again motivated by her values—directs top-down attention to favor the color naming response. Nonetheless, the reading response wins out in the decisional competition due to the effects of various stochastic factors. The incorrect response that occurs thus flows from *other* features of her psyche; her values in fact "push against" the response.[22]

We can home in further on what exactly allows valuationist views to avoid the fallibility paradox while reasons-responsiveness theorists fall prey to it. The key difference is that valuationist views can make fine-grained distinctions where the reasons-responsiveness cannot.

Consider an agent *S* performing a series of actions that arise from a set of psychological processes *P*. Call the set of these actions *A*, and further suppose that *P* has some low rate of noise-based errors. As we have seen, reasons-responsiveness views employ a threshold to assess *S*'s moral responsibility for the elements of *A*. If *P* meets the threshold, then *S* is morally responsible for *every* action in *A*; if *P* fails to meet this threshold, then *S* is not morally responsible for *any* element of *A*. That is, so long as it is *P* that is operative in producing the actions, it is not possible for a reasons-responsiveness view to say *S* is morally responsible for some of the actions in *A* and not morally responsible for others.[23] This is what I refer to as the "course-grain" constraint on reasons-responsiveness views: reasons-responsiveness views must treat all elements of *A* identically.

It is worth emphasizing that, for the reasons that I noted earlier (see Section V.C), the problem of course-grainedness isn't specific to *mechanism*-based versions of reasons-responsiveness views. There I noted that the most natural way to understand *agent*-based views is that they say the relevant *P* is not a subset of the agent's psychological processes, but rather is *all* the psychological processes of the agent. Having selected an expansive *P*, agent-based views still face the exact same issue: they must next place a threshold for when *P* is sufficiently responsive for moral responsibility. Thus they too have the problem of being fundamentally

[22] Earlier I discussed the "different mechanism" strategy that might be taken up by reasons-responsiveness theorists. Some readers of that section might have thought about the following strategy for mechanism individuation: In the 996 trials in which Fei performs the word reading response, the mechanism that issues in action produces an action that is appropriately caused by and expresses Fei's goals (to read the word rather than say the ink color) and values (caring for animals). On the four trials in which she makes an error, the mechanism that issues in action produces an action that is not caused by, and in fact conflicts with, these goals and values. Could this difference be a *principled* basis, one that has strong roots in intuition, for saying that there are different mechanisms at work on the 996 success trials versus the four error trials? The discussion in the present section serves to show why this strategy is misguided. This approach to mechanism individuation is sufficiently different from the standard reasons-responsiveness approach, and sufficiently similar to the valuationist approach, that a reasons-responsiveness theorist who takes this tack is essentially collapsing his view into a form of valuationism (see McKenna, "Reasons-Responsiveness, Agents, and Mechanisms" for further discussion of related points).

[23] Here, as I have been doing throughout this essay (unless explicitly noted otherwise), I am assuming all other conditions for moral responsibility are met.

course-grained: given their selection of *P*, they too must treat all elements of *A* identically.

Valuationist theories, in contrast, don't rely on arbitrary thresholds on the responsivity of *P*, and this enables them to make the critical fine-grained distinctions needed to address the fallibility paradox. Even when *P* is operative across a large number of cases, valuationist views say the agent is morally responsible for the relevant actions in all and only those cases where *P* produces an action that expresses the agent's evaluative point of view. In the rare instances in which *P* produces an action that does *not* express the agent's evaluative point of view (for example, due to noise-based errors), the agent is *not* morally responsible for the action. Thus we are able to say the agent is morally responsible for most elements of *A*, but not all—precisely what is needed to avoid the fallibility paradox.

The fallibility paradox, then, isn't just a counterexample to reasons-responsiveness views of moral responsibility. It is also the basis for a strong argument in favor of valuationist views. Reasons-responsiveness views are inherently course-grained, but the fallibility paradox highlights that our intuitions about moral responsibility don't follow a course-grained pattern. Valuationist views, in contrast, are extremely well positioned, perhaps even uniquely positioned, to capture the fine-grained pattern with which we ordinarily attribute moral responsibility.

Philosophy, Psychiatry, University of Michigan