The Monist, 2020, 103, 294–311
doi: 10.1093/monist/onaa005
Article

Responsibility, Reflection, and Rational Ability

Dana Kay Nelkin*

ABSTRACT

This paper takes as its starting point the thesis that one is responsible for one's actions insofar as one has the ability to act for good reasons. Such a view faces a challenge: it is plausible that only beings with the ability to *reflect* are responsible agents, and yet it seems that not only is it possible to act for reasons without reflecting, it seems to happen quite frequently. Thus, advocates of the rational-ability view of responsibility must either reject as a necessary condition that responsible agents must have the ability to reflect, or locate a plausible role for reflective ability. In this paper, I propose and assess a variety of ways to meet this challenge.

1. INTRODUCTION: A CHALLENGE FOR RATIONAL-ABILITY ACCOUNTS OF FREE AND RESPONSIBLE AGENCY

A number of theorists who otherwise disagree on important issues converge on the idea that being able to act for good reasons is the key to free and morally responsible agency. This idea takes many different forms, and it can be combined with a commitment to either the compatibility or the incompatibility of such agency with determinism, or even with agnosticism on that point. While some versions of the view add a further condition of some sort, on many other versions, being free and responsible on a given occasion just *is* to have acted with the ability—or a sufficiently strong ability—to respond to the good reasons there are for acting in a certain way. In other words, it is both necessary and sufficient for acting freely and responsibly—and so, depending on whether one acts well or badly, blameworthily or praiseworthily—that one had the (sufficiently strong) ability to respond to reasons at the time. Call these "rational-ability" views.[1]

Rational-ability views are generally proposed as offering conditions for a robust notion of moral responsibility, as opposed to weaker notions such as causal responsibility. In what follows, I will assume that what is at stake is a notion of responsibility as accountability, so that responsible agents are accountable for their actions, which includes being liable to negative setbacks of their interests when they act badly.[2] Thus, much is at stake in identifying the conditions of responsible agency, and rational-ability views propose a single unifying condition as both necessary and sufficient for it. At the same time, being able to act for the reasons there are in a given

*University of California, San Diego

situation may in turn require a number of other abilities, including perceptual, cognitive, emotional, and motivational ones. These various abilities are linked together in that they collectively provide agents with the ability to respond to reasons.

Despite—or perhaps because of—its ultimate simplicity, this kind of view generates an interesting puzzle that arises as a result of two considerations taken together. The first is that the ability to reflect seems essential to responsible agency. One way this commitment arises is via a commonly recognized *desideratum* of a good theory of free and responsible agency, namely, that it should explain why such agency is special to *persons*. As Harry Frankfurt highlights in his highly influential "Free Will and the Concept of a Person," free agency is something we take to categorize persons, as opposed to, say, squirrels (1971). A good theory of free agency ought to explain this fact. And what distinguishes persons from squirrels (so far as we know)? We are *reflective beings*; we have the ability to reflect on our own reasons. In that well-known paper, Frankfurt proposes a condition on free agency that an agent act freely when she acts on a first-order desire that she desires (at the second order) to be effective in action, which is naturally taken to be a reflective ability. To have the ability to have such second-order desires could both explain why only persons are candidates for free will and why we attribute free action when we do. But others—arguably including Frankfurt on later occasions—take the kind of reflective ability at issue to go beyond the mere possibility of possessing passive second-order attitudes. They take it to be a kind of goal-directed activity on the part of agents. For example, Christine Korsgaard describes a kind of reflective ability in the following terms:

> What this means is that the space of reflective distance presents us with both the possibility and the necessity of exerting a kind of control over our beliefs and actions that the other animals do not have. We are, or can be, active, self-directing, with respect to our beliefs and actions to a greater extent than the other animals are, for we can accept or reject the grounds of belief and action that perception and desire offer to us. We can actively participate in giving shape both to the conception of the world in light of which we act and to the motives on the basis of which we act and ultimately, in both ways, in giving shape to ourselves. And it is the same fact that we now both can have, and absolutely require, reasons to believe and act as we do. (Korsgaard 2009, 32)

Or, relatedly, one might take reflection to be assessment with a further goal in mind, namely, the adopting of reasons in the service of deciding what to do on the basis of those reasons.[3] This might be best described as "rational deliberation," though it, too, often goes by the label "reflection." Thus, we have more and less robust notions of reflection ranging from a notion of reflection as the possession of second-order pro-attitudes that take first-order attitudes as objects to a notion of reflection as an active goal-directed activity of assessment with the goal of adopting good reasons for action.[4] As we will soon see, a puzzle arises for just about *any* interpretation of the reflective ability that distinguishes us as persons who are uniquely possessed of free and responsible agency when combined with a second consideration.

The second consideration has come to the fore as philosophers have voraciously consumed a large body of empirical psychological results that suggest that the vast majority of human actions is not performed with reflection on our reasons, and, further, without the ability to reflect on our reasons most of the time.[5] This is bolstered by evidence that when we are explicitly asked to reflect on the reasons for which we acted, we often fail to accurately represent our beliefs and motives. To take just one dramatic study as an 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.[6] Putting this together with empirical evidence that suggests we are often in a kind of "flow" in acting—whether driving or playing music or writing or playing sports—that doesn't seem to leave time or room for reflection. Where we are unable to articulate our reasons very well after the fact, it seems that we very often act without reflection, and we sometimes perform better when we don't reflect.[7] And yet, it would be revisionary in the extreme to think that almost nothing we do is done for reasons. Thus, it appears that the ability to act for reasons on any given occasion does not entail that one reflect on that occasion. But if we *can* act for reasons without *actual* reflection, it isn't clear that we would need the *ability* to reflect. Without such an explanation, it appears that the ability to act for reasons does not entail an ability for reflection. And I will try to show in what follows that there is no obvious explanation. In contrast, in some cases, it seems that acting in a certain way implicates an ability that we do not exercise. Someone who runs a mile in 4 minutes would seem to have the ability to run a mile in *over* 4 minutes, or someone who completes a mathematical proof implicates, though may not exercise, the ability to do simple addition. In these cases, we can explain why one would have to have the ability to run a slower mile or do simple addition. But, looking ahead, the case at hand is importantly disanalogous from these, and it is not obvious what the explanation is for why an ability to reflect would be required by acting on reasons.

In addition to empirical support for this idea, a recent conceptual line of reasoning against the entailment has also been influential. According to a so-called "regress" argument, reflection itself is an activity done for reasons, and yet, if it required an act of reflection for its reasons-responsive quality, then we would be at the start of a regress of infinite acts of reflection.[8] Thus, it must be possible to act for reasons without the *possibility* of reflecting. Again, this is not yet to say that we can act for reasons without having the *ability* to reflect; but considerations of parsimony suggest that to avoid this conclusion, we would need an explanation of why having such an ability is necessary for acting for reasons, and as I will try to show, it is not obvious what the explanation is. As Nomy Arpaly and Tim Schroeder, advocates of the regress argument, conclude, while there might be a role for reflection in our lives, it is "contingent, intermittent, and modest" (2012, 209). So we have at least two routes to the second important consideration, namely, that rational agency with its various rational abilities does not require reflective capacity. This second consideration has led to a lively and productive debate about the nature of rational agency, some of which I will have occasion to turn to in section 4. But even if we were to accept that

robust rational agency does not require reflective capacity, noting this second consideration is enough to generate a serious puzzle for the view that a necessary and sufficient condition on acting freely and responsibly is having an ability to respond to reasons when one acts.

For if possessing such an ability is necessary and sufficient for being responsible for an action (or omission), as the rational-abilities view suggests, and satisfying that condition does not obviously implicate any sort of reflective ability, then it is not clear how we can embrace the idea that responsible agency *also* requires reflective ability. How can we explain how the rational-ability condition and the reflective-ability condition are both true? It will be helpful to set out the puzzle in its boldest form so that we can most easily see the explanatory task facing the rational-ability view. In slogans: rational ability entails responsible agency; responsible agency, as the first consideration spells out, entails reflective ability; thus, rational ability entails reflective ability; but, by the second consideration, rational ability does not entail reflective ability. Call this "The Puzzle" faced by rational-ability views.

Of course, we could simply reject one of the claims that together create The Puzzle. But it is also possible to avoid inconsistency by instead disambiguating different interpretations of "ability" in the claims that make up The Puzzle, and the main project of this paper is to explore this possibility. In order to see whether and, if so, how, we can achieve this, we first need to articulate more precisely just what sort of ability is at issue in the most plausible rational-ability view. We also need to articulate more precisely just what sort of reflective ability is implicated by responsible agency. I take up each of these tasks in sections 2 and 3 respectively. At this point, we will have narrowed down possible ways to respond to The Puzzle by reading its claims in a way that make them perfectly consistent. But important work will still be left to do, as even if it is possible to find a role for reflective ability consistent with the rational-ability view, it still remains to *explain* how the different abilities—rational and reflective—are related. As I will argue, there is no obvious way to do this. I take up this task in section 4. In section 5, I conclude with a brief comparison to alternative responses to The Puzzle that include rejection of one or more of its claims.[9]

2. THE OPPORTUNITY ABILITY TO RESPOND TO REASONS

In what follows, I set out the rational-ability view I favor, which I call the Quality of Opportunity view, and then contrast it briefly with some influential competitors along a variety of dimensions. We can begin by identifying just why so much is stake when it comes to figuring out the nature of responsible agency. Among other things, being responsible agents opens us up to being held accountable, and to being appropriately blamed and praised depending on whether we act badly or well. In order to be accountable, it must be appropriate for others to hold one to moral demands. In turn, for this to be appropriate, one must have the ability to comply with them. Further, given an understanding of the contents of the relevant demands as demands to act (or not act) in certain ways for the right reasons, it seems that the ability to so act (in the right ways, for the right reasons) is required in order for demands to be appropriate.[10] In particular, one must have the opportunity to avoid wrongdoing. Importantly, it is not enough to have what is sometimes known as a "general

ability"—a general skill or a kind of know-how—in order for a demand to be apt. Something seems to have gone wrong if I demand that someone stop an assailant if she is tied up and has been given a drug that inhibits her cognitive processing, even if she has the highest level of training and skill in chasing down assailants. To have an opportunity to avoid wrongdoing, one must have some general competence, but it is also the case that one must not be prevented in a relevant sense from *exercising* one's general skills and competence on the particular occasion.

Going a step further, in order to be *blameworthy* in the accountability sense, one must not only have an opportunity to avoid wrongdoing, but also have a *high enough quality* opportunity. These satisfaction conditions on blameworthy action in the accountability sense gain support from the fact that a conception of blameworthiness in terms of the fair opportunity to avoid wrongdoing best captures the wide variety of commonly recognized excuses. For example, it explains why we recognize excusing conditions ranging from ones that compromise our normative competence (via either cognitive or volitional impairments) to ones that impose situational constraints such as conditions constituting duress. What ultimately brings these together is that both such impairments and such constraints can in their own ways result in a lack of opportunity (or a lack of a high-enough quality one) to avoid acting wrongly. Working backwards from excuse to blameworthiness in the accountability sense, we can see that the latter is then instantiated just when one acts wrongly and at the same time possesses an opportunity of high enough quality to avoid wrongdoing, or, more positively, to do the right things for the right reasons.[11]

Importantly, the quality of opportunities is scalar. The worse an opportunity is, say, because due to its features it makes it very difficult to act rightly, the more mitigating of blameworthiness it is if one fails to do so. On the flip side, if one does act well for good reasons in such circumstances, one is more praiseworthy than if one acted well when it is easy to do so.

Before going further, two clarifications will be helpful. First, to avoid confusion, it is worth noting that some writers, including influential legal theorists, reserve "opportunity" to refer to situational factors alone.[12] But because situational factors and general competence are not *independent* factors in determining whether one *can* in the relevant sense, avoid wrongdoing, it is important to have a single label for the relevant ability. In what follows, then, I will use "opportunity ability" to capture the notion of ability at issue that is a function of both one's skills and competence on the one hand, and the congeniality of the situation on the other. This encompasses what others have called "opportunity."

Second, at this point, a number of further questions arise. Is having an opportunity to act differently than one actually acts dependent on the truth of indeterminism? Or is it dependent in a different way on the nature of the laws of nature, such as whether they are Humean or necessitarian?[13] These are deep and difficult questions. For present purposes, however, I believe that it is possible to remain neutral on these questions. The reason is that The Puzzle applies to opportunity-ability views regardless of how they answer these questions.

Now, in grounding responsibility in the nature of one's opportunities to meet relevant demands (and ultimately in the obligations or standards on which these rest),

the Quality of Opportunity view is best categorized as a "control" view, in contrast to views that emphasize the nature of evaluative judgment or quality of will expressed in action or attitude. It is essential that one have control in order to be responsible, in virtue of one having sufficiently good opportunities to act well.

There is one more piece to be added to the view. Along with many other control theorists, I add to the theory by adopting a so-called "tracing" condition, in order to accommodate situations in which intuitively it appears that one lacks an opportunity at the moment to act otherwise, such as cases of unwitting omissions like forgetting to keep a promise. In other words, even if one doesn't meet the conditions for control (or, more specifically, on the account I've set out, one lacks an opportunity of high enough quality at the time of the act or omission), one might be responsible for that act or omission. One is responsible for that act or omission in virtue of an earlier moment in which one did meet the relevant conditions and at which the risk of the later act or omission was foreseeable. Thus, perhaps one earlier chose not to set an alert on one's phone to remind one to keep one's promise, or the thought occurred to one that one could do so and just let the moment pass; in either case—choice or mere opportunity—one is responsible for the omission in virtue of what one did or did not do with the earlier opportunity.[14]

Combining these elements, then, the account of responsibility I propose here is one that takes the quality of opportunity to be central. One can be responsible and blameworthy if one has a high enough quality of opportunity and does not take it at the time of action (or omission), *or* if one earlier had a high enough quality of opportunity at which time one was aware of the risk of acting (or omitting) badly later.

While this is an incomplete sketch of the proposal, it will help to contrast it briefly with some alternatives on key dimensions. First, theorists who focus on rational ability divide on whether it is the ability of the agent or the mechanism on which the agent acts that is central.[15] Second, the proposal at hand is so far silent about its implications for the truth of particular counterfactuals or descriptions of alternate possible worlds, whereas at least some related accounts either seem to understand the relevant rational abilities entirely in terms of the truth of relevant counterfactuals, or at least to take there to be a necessary entailment. Fischer and Ravizza (1998) can be read as taking the kind of control in question to bottom out in the responses to reasons that the mechanism on which the agent actually acts *would* provide in sufficiently many similar possible worlds and in an intelligible pattern. Brink (2013), adopting an agent-centered rational-ability view, also understands the ability in question in terms of counterfactuals. For example, in determining whether a seminarian rushing by a person in need had a sufficiently strong ability to have discerned and then acted on the reasons to stop, we should look at nearby worlds (such as ones in which she was not in quite such a rush) to see if she stops. If she would stop in sufficiently many such worlds, we can conclude that she has the ability in the actual world.[16]

Finally, rational-ability views can differ on how they measure degrees of responsibility. On the view put forward here, degrees of responsibility track aspects of the quality of opportunity, such as the degree of difficulty in doing the right or good thing. But at least some views that appeal to possible worlds lend themselves to

cashing out degrees of responsibility in terms of distance of relevant possible worlds in which agents respond to reasons.[17]

As we will see, these difference in views that feature rational ability—mechanisms vs. agents, roles for counterfactuals and possible worlds, and understanding of the measure of degrees of responsibility—all provide the views with distinctive resources with which to address The Puzzle. Before we can do so, we first need to turn to the other ability in question, namely, reflective ability. In the next section, I set out several candidates for how to understand this ability.

3. CANDIDATE CONCEPTIONS OF THE REFLECTIVE-ABILITY CONDITION

The present reflective opportunity-ability condition

On one conception of the reflective ability needed to be responsible for an action or omission, an agent must have the opportunity ability to reflect at the time of action or omission. That is, if Allie is responsible (and, perhaps blameworthy) for her lying at t1 to a reporter at a press conference, then it must be the case that she has the opportunity at t1 to reflect on the reasons for lying and for not lying. This seems too strong a requirement for a number of reasons, even when we take into account the intuitive pull of the need for reflection. One way of bringing out the point is to imagine that Allie had reflected on her reasons earlier, and then committed to lying, having trained herself to do so without inhibition based on certain cues. In that case, the fact that in the moments just prior to and during the telling of her lie she can't now reflect is no bar to properly holding her responsible for the lying. This is thus not a plausible candidate for the reflection condition on responsibility. More generally, we are reluctant to withdraw our widespread attributions of responsibility when faced with the kind of empirical evidence with which we began, supporting the conclusion that reflection is relatively rare.

The general reflective-ability condition

On a second proposal, the reflective ability in question is just a general ability or competence. What one needs is some skill or set of skills or competence, and one can possess this without having an opportunity to actually exercise one's skill on a given occasion. Allie might have the general ability to tell the truth, but at a particular moment, having been hypnotized to lie, say, there is an important sense in which she cannot exercise it. This proposal has a number of advantages. First, at least sometimes, the way that the reflective ability is presented is consistent with its being a claim of a general competence. Responsible agents are *reflective beings*. Such a claim does not suggest that to be responsible for a particular action or omission, one must have an opportunity ability to reflect on one's reasons on that particular occasion or with respect to one's particular reasons for that action or omission. Second, it fits nicely with the empirical evidence that we do not reflect often, and it does not provide a start to a regress argument. Third, and especially importantly for present purposes, if it were correct, it would allow us to see that The Puzzle is based on an

equivocation between the general ability to reflect and the opportunity ability to reflect.

At the same time, it is not enough to point out an equivocation in the claims of The Puzzle, for an explanatory burden remains to be discharged. In particular, we need an explanation of why the possession of an opportunity ability to respond to reasons would entail the possession of a general ability to reflect on reasons. Consider an analogy: if I reported that my having a current opportunity ability to run a five-minute mile right now entails that I have the general ability to run a four-minute mile, that would seem to call out for explanation. (In fact, it would seem implausible barring an explanation.) Similarly, we would need some explanation for why the opportunity ability to respond to reasons would entail a general ability to reflect on one's reasons. I return to this suggestion in the next section. But first, it remains to consider additional candidates.

The previously possessed opportunity ability (tracing) condition

On a third proposal, we can locate a parallel to the opportunity-ability proposal for rational ability with its built-in tracing option. On this proposal, one must have the opportunity ability to reflect, either at the time of action or at an earlier time appropriately related to the later time, as before.[18] This proposal avoids conflict with empirical results showing the rarity of reflection. At the same time, when combined with the Quality of Opportunity view of responsible agency, an explanatory demand immediately arises: why should reflective ability of this sort be entailed by rational ability? We will take up this explanatory question in section 4.

In sum, we have at least two remaining contenders for the Reflective-Ability Condition that are consistent with the Rational-Ability Condition. By appealing to either the general ability to reflect or to a tracing rider on the opportunity ability, we can point to an equivocation in The Puzzle.

4. TURNING TO THE EXPLANATORY TASK: HOW IS THE REFLECTIVE ABILITY RELATED TO THE RATIONAL ABILITY?

So far, so good, but important work remains. For, as we also saw, it is not clear *why* we should need the general reflective ability if we don't need opportunities to use it on a wide variety of occasions when we act perfectly responsibly. We need an explanation.

I see at least three possibilities. The first appeals to a notion of reflection that is more like the relatively thin notion Frankfurt put forward, namely, the possession of second-order attitudes toward first-order ones. Particularly when it comes to moral reasons (though not necessarily exclusively), we can only access such reasons by means of concepts that take first-order attitudes as objects. For example, to understand other people's interests or rights as reasons, we must understand that they have desires or ends or affective states, and be able to prioritize our promotion of some over others. Arguably, to track others' interests in the right way as reasons-giving, even in a nonreflective way, we must have *concepts* that presuppose a kind of second-order awareness and capacity to value first-order states. To respond to truly

moral reasons, then, requires reflective capacity of at least this kind even if one can act morally without the opportunity to reflect at a given moment. In particular, one must have exercised actual opportunity abilities to reflect in a way that allows one access to moral reasons in order to develop relevant moral concepts.[19]

A second, related approach is to distinguish between two conceptions of acting for reasons. As Agnieszka Jaworska (2016) explains the distinction, there is a weaker and a stronger interpretation of "acting for a reason." On the weaker interpretation, "one's reason for action is simply a consideration one takes as speaking in some way in favor of the action," whereas on the stronger interpretation, one acts under the "guise of the good" or, in other words, "one evaluates the action as having a good-making feature" (69).[20] In taking one's action to have a good-making feature, one "believes that one is correct" and implicitly recognizes that one might be mistaken. This second way of acting for a reason requires at least a second-order attitude of sorts, namely, to take one's own attitudes to be correct or to take oneself to be getting things right. If the opportunity ability to act for reasons incorporates the stronger interpretation of "acting for reasons," then we have a different route to a vindication of an ability for reflection in the thinner sense. In fact, it appears that the very opportunity ability to act for reasons will itself entail an opportunity ability to reflect in this thinner sense.[21]

Both of these approaches provides vindication for what seems to be a thinner notion of reflective ability than ones like *deliberative* ability. Take the first approach: reflection in the form of higher order attitudes is required in order to acquire *concepts* that one must have in order to respond to certain kinds of reasons involving rights and interests of others. This is no doubt important, and arguably it sets persons aside from other beings. But it seems somewhat far from the "reflective distance" that is supposed to give us some added control over our own mental states that Korsgaard mentions. The second approach comes closer than the first in that it requires an evaluative understanding of our own reasons, and the ability to reflect in this sense is implicated in the very opportunity ability to act on reasons. But we would need an argument for interpreting the rational-abilities view in this stronger way, and, in any case, it is not obvious that the ability to see yourself as possibly mistaken in your evaluations gives rise to an ability to actively deliberate. Is there a way to implicate a more robust notion of reflective ability that takes reflection to be an active exercise of an ability?

Arguably, such an active ability to reflect is systematically in place for responsible agents. Given our fallible nature, it seems that reflection is the natural corrective to our getting things wrong on any given occasion. If we have unreflectively latched on to bad reasons for acting, then the last line of defense for providing us an opportunity to correct course would seem to be an opportunity ability to reflect on the reasons there are. It might be possible for the world to somehow cooperate so that before we are otherwise about to act badly we get a situational adjustment. Perhaps we are lucky enough to live in a world where we get correctives simply through association—the right song comes on the radio at the right time, giving us a subliminal message, and we end up being sensitive to a reason we would not otherwise

recognize, say. But the systematic way to have opportunities for correction comes from a general ability to reflect, with the opportunity to exercise them when it is needed.[22]

This is to appeal to something like the opportunity ability with a tracing rider. Note, however, that it is not necessary in every case of responsible agency that we have the possibility of correction. When we get it right, we do not need to have had a corrective outlet. And note that we need not have distinct opportunity abilities indexed individually to every action we perform. We can have opportunities to change course in more general ways that can leave us on the hook—or, more positively, liberate us and allow us to celebrate—and that cover a whole set of our acts and omissions.[23] Still, the fact that an opportunity ability is necessary in some cases shows that the general ability is necessary, since to have the opportunity ability itself requires a general ability.[24]

But at this point we can ask whether even such a general ability is necessary for responsibility. Take the case of Mark Twain's Huck Finn who does the right thing, and arguably for the right reasons, but upon reflection thinks he is making a terrible mistake.[25] Many have the intuition that Huck is not only responsible but praiseworthy for his action. Does having the general ability to reflect help him achieve praiseworthiness here? It is not at all obvious that it is a help, rather than a potential hindrance.

In reply, note that there are two possible Huck Finns to consider here. One is like us—sometimes he gets things right with or without reflection, and sometimes he doesn't. For this Huck Finn, at least in many cases, it seems that the opportunity to reflect will be essential to make responsible agency possible, and certainly a general ability is needed in that case if one is to have the opportunity ability to employ one's competence. But another Huck Finn is someone who just always get things right for the right reasons (and perhaps would have a worse record had he used reflection). Perhaps for such a person, who is in one way Godlike, such an ability would not be necessary.[26]

Recall Arpaly and Schroeder's conclusion that the role of reflection is "contingent, intermittent and modest." They were concerned primarily with the role of reflection in acting for reasons. When we transpose their conclusion to responsible agency, the concession just made vindicates the parallel idea that reflection is indeed contingent for responsible agency. It is also correct that it is intermittent in the sense that actual reflection happens infrequently. But on this proposal what we need is the opportunity to reflect, and while even such opportunities might be few and far between, if they might in a crucial way account for a large number of actions and omissions, then "intermittent" might be a misleading term.[27] And most importantly, it does not accept that the role of reflection is modest. For the ability must be in place for fallible but largely responsible beings like us. The difference in our judgments of modesty might come down to a difference in standards for modesty; but it might also be that in the context of *responsible* agency, the need for reflective ability as a systematic corrective takes on added importance.

The question now facing us is whether this immodesty is enough to capture the initial appeal of the requirement of reflective ability for responsible agency, given

that we have conceded its contingency. At the very least, it is not "merely" contingent. Its role is neither practically dispensable nor perhaps even conceptually dispensable for beings who go wrong if we are in need of a systematic path to correction. Thus, while reflective ability is not essential to responsible agency on this view, it is essential to responsible agency in fallible beings. It is the only systematic way of providing us the opportunities we need to be responsible agents in the face of fallibility.

In sum, there are at least three candidate ways of understanding reflection so as to see rational ability as requiring reflective ability, in increasing order of robustness. The more robust the reflective ability, the more promissory notes are required to defend the entailment.

5. THE ALTERNATIVES: REJECTING EITHER THE RATIONAL-ABILITY CONDITION OR THE REFLECTIVE-ABILITY REQUIREMENT

It remains to briefly consider and compare alternative responses to The Puzzle that reject one or more of its claims. A full assessment of the alternatives would require an in-depth examination of the particular positive arguments for the two considerations that, together with the rational-ability view, create The Puzzle. Instead, here I focus on some specific ways of deploying the resources identified earlier to support one or another alternative. Interestingly, we can now see how different versions of the rational-ability condition and the reflective-ability requirement can also affect the quality of these alternatives.

Reject the reflective-ability requirement

This option has us rejecting the idea that reflection is essential for responsible agency. The empirical results purporting to show how rarely we reflect on our reasons, despite getting around quite well in the world, doing things we take ourselves to be responsible for, certainly motivate such a view. But the idea that reflection—particularly the general ability to reflect—is necessary for being responsible agents in the first place is a powerful one, and, as we have seen, it is not undermined by these empirical results that show the rarity of reflection.

One way of rejecting this requirement, while offering at least a partial explanation of the *appeal* of reflection, is to see reflective ability as something that can make our rational ability *better*. One can combine this with the thesis that reflection is a normative ideal.[28] Vargas puts the point this way:

> Perhaps the most promising strategy might be to emphasize the ideality of accurate reflective agency, but to concede that in ordinary life, the most we can hope for is some (relatively low) threshold of responsiveness to reasons. Such a theorist might yet hold that we are rational enough (e.g., sometimes in conscious, deliberate ways, sometimes not) to make sense of responsibility and agency if, for example, in the ordinary case agents would (given full

> information) see themselves as responding to reasons enough for acting the way in which they did. (2018, 261)

This is an interesting suggestion. The idea seems to be that while a lack of reflection might decrease the level of reasons-responsiveness or ability to act for reasons, it need not decrease it beyond a threshold for responsibility.[29]

How does reflective ability make rational ability better? At least one way to see this is to return to the suggestion (put forward by Vargas, among others) that we can understand that latter ability in terms of the proportion of nearby possible worlds in which one does respond to reasons, even if in the actual world one does not. If one counts as *more* reasons-responsive, or as having a stronger ability to act well if this is the case, and having reflective ability, even unexercised in the actual world, makes it the case that one *would* have acted on good reasons in a greater number of possible worlds, then one would count as having a stronger ability to act well in the actual world if one has the reflective ability.

There is some reason to doubt, however, that this is really the fundamental measure of how good one's ability is. Simply because someone *would not* act well in a wide range of nearby possible worlds does not mean that they could not.[30] Perhaps there are additional ways to see how a particular rational opportunity ability improves just in virtue of having reflective opportunity ability, while still thinking that our ability is "good enough" (to use a phrase of Fischer's [2017]) without it. And perhaps this is the best that we can do. But given the power of the initial intuition that reflective ability is essential to responsible agency, if it is possible to find a way of retaining it that is explanatorily satisfying in the face of such results, we would have good reason to do so.[31]

Reject the rational-ability view

We could instead reject the rational-ability view. Of course, one way to do so is to reject both its claim to the necessity and its claim to the sufficiency of the rational-ability condition. This would be the more radical move. There is extensive debate on this point, and I do not have the space to enter it in any substantial way here. But it is worth noting that one of its main competitors—the so-called "quality of will view" which takes our responsible agency to rest on our actions and omissions revealing some quality of our wills—faces a parallel puzzle to The Puzzle. It appears that one can reveal a bad or malicious quality of will, say, without having reflected, or without the opportunity ability to reflect. So simply rejecting the rational-ability view altogether in this way brings its own puzzle with it.

A less radical way of rejecting the view is to retain the claim that rational ability is necessary for responsible agency, while rejecting the claim that it is sufficient. One way to accomplish this would be to invoke reflection in a further requirement that one must in some way have *taken* responsibility for one's actions; another would be to invoke reflection in a further requirement that one's responsible actions must be *one's own* in some important sense.

The prominent account of Fischer and Ravizza (1998) takes the first way. On their view, a necessary condition for responsibility is a reasons-responsiveness one

that does not itself entail reflection. But, on their view, it must be combined with a so-called "historical" condition to provide a sufficient condition for responsibility. That condition requires that one have—at some time in the past—taken responsibility for the reasons-responsive mechanism on which one acts. And, in turn, this requires our having accepted that acting on such a mechanism makes us the appropriate target of reactive attitudes for what results from it (Fischer and Ravizza 1998). Now this suggestion has been criticized on various grounds, including that there is no obvious and stable understanding of "mechanism" here, so that it becomes implausible that we could take responsibility in the relevant way for our mechanisms when we don't even have an intuitive understanding of how to individuate them.[32] These criticisms appear difficult to answer. But one might consider a variant of the suggestion that does not rely on mechanisms. For example: to be responsible for any action, one must have taken responsibility for one's actions performed with rational ability as a whole by seeing oneself quite generally as a responsible agent. Interestingly, this would be to appeal to *a* kind—and at least a moment of—reflection on the part of responsible agents.

A serious difficulty facing this suggestion, however, is that we do not have an explanation for why *taking* responsibility is essential for being responsible. After all, we often hold people responsible for things for which they explicitly disavow responsibility, or for unwitting omissions; the fact that they have not taken responsibility is not generally taken to be an excuse, and so this suggestion appears unmotivated here.[33]

Turn then to a second way of incorporating reflective ability into a distinct necessary condition for responsible agency. It is to argue that (i) actions are only really *one's own* when they are the objects of reflection and endorsement, and (ii) we are only responsible for actions that are our own. Perhaps this is close to the way that Frankfurt originally thought about the situation—though he seemed to think this could serve as a sufficient condition without even a rational ability needed.[34] Now (i) is even stronger than the requirement of a reflective *ability*, in requiring actual reflection. And this has seemed to many to be too strong, even when trying to capture some notion of an action being *one's own* in some important way. For example, Gary Watson points out that one might be "fully behind" one's driving at one hundred miles per hour on the freeway, despite not reflectively endorsing such an activity (1975). Similarly, Huck Finn would seem to be a counterexample to the conjunction of (i) and (ii). While there is a rich and interesting debate about how exactly to capture the notion(s) of an actions being "one's own," it is not at all obvious how answering it could vindicate the idea that a separate reflective-ability condition should be placed on responsible agency.

I do not mean to say that there is no separate and additional necessary condition on responsibility beyond rational ability that invokes reflective ability. And it may be that this really is the only option that fully captures the appeal of the reflective-ability requirement. But it is as yet unclear exactly what such a condition would be and exactly why it should be required for responsible agency.

6. CONCLUSION

I have presented a puzzle for rational-abilities views of responsibility. The ultimate task is to explain how acting with rational ability can be both necessary and sufficient for morally responsible agency while at the same time accounting for the strong appeal of the idea that it is only beings with reflective ability that can be responsible agents. I have set out the beginnings of what I take to be promising ways in which reflective ability might be connected to the rational abilities that ground our responsible agency, and at the same time, set out reasons for thinking that there is more work to be done in choosing among them.[35]

NOTES

1. Despite interesting differences among them, examples of proponents of rational-ability views include Brink, Nelkin, Vargas, Vihvelin, Wallace, and Wolf. Fischer and Ravizza (1998) is a central and very influential text that features reasons-responsiveness as a key condition for responsible agency, but it also recognizes an additional condition, notably a necessary historical condition that one has in the past taken responsibility for acts that result from one's reasons-responsive mechanisms. I discuss this suggestion in section 4, but because of its inclusion do not count Fischer and Ravizza's view as a (pure) rational-ability view. McKenna (2000, 2012) criticizes their historical condition, and in correspondence he says he is also open to a nonhistorical version of such a view and, in that case, could also count as an adherent of a rational-ability view as I have defined it.
2. See Watson (1996) who contrasts this notion with responsibility as attributability. Elsewhere I have argued that being responsible in the accountability sense is equivalent to being responsible in a way that makes one open to *desert*, as well. See Pereboom (2014) for an understanding of basic desert, and Nelkin (2016) for an argument for an equivalence. Importantly, I do not accept that people's being deserving of a certain treatment by itself provides reason for others to give them what they deserve; however, being so deserving does make one liable to receiving it under proper conditions that include its being necessary for a significant good.
3. See, for example, Nelkin (2011).
4. There is an interesting and related debate about whether *consciousness of salient considerations* is required for responsibility. See, e.g., Levy (2014) for an extended argument in favor, and Carruthers and King (forthcoming) for a survey and a defense of the opposing view.
5. See, for example, Doris (2016) for a survey of empirical literature supporting this conclusion.
6. See Bateson et al. (2006). See Doris (2016, 41–64) for discussion of this and many related studies. This does not by itself show that they were unaware of the *reasons* on which they acted, but it does cast some doubt on the idea that they could correctly explain why they acted as they did.
7. See Railton (2009) for one prominent philosophical discussion of research and implications of the "flow" literature. Railton points out that actions done in the "flow" can include ones with morally relevant qualities. For example, a driver in a hurry might in an instant slow her car and wave to what she has realized is an elderly driver rather than speeding past him, causing him to relax into a smile (Railton 2009, 31). But see Montero (2016) for an important re-evaluation of the empirical evidence supporting the idea that experts in particular, such as expert dancers, chess players, or mathematicians, do not reflect when performing at a high level, arguing against a "Just Do It" principle and in favor of a "Cognition in Action" principle. According to that principle, when experts are having (near) optimal performances, they "frequently employ some of the following conscious mental processes: self-reflective thinking, planning, predicting, deliberation, attention to or monitoring of their actions, conceptualizing their actions, conscious control, trying, effort, having a sense of the self, or acting for a reason" (Montero 2016, 38). As Montero suggests, given that this principle invokes a frequency claim, it is not inconsistent with a view that suggests that often we act without any of those conscious mental processes, and we can see the two views as different in emphasis (50). Julia Annas (2011) presents a picture of skilled action and expertise (including moral expertise) that may not require reflection on the moment, but which does require past episodes of reflection on reasons that "leave a trace," an idea to which I will return in section 4.
8. See Arpaly and Schroeder (2012) and Railton (2009). See Kornblith (2012) for additional arguments that reflection is not necessary for reasons-responsiveness in belief or action.

9. Others have set out related challenges. For example, Doris (2016) poses an extended challenge for any theory that requires reflection for responsibility, and takes it to succeed in showing that reflection is not in fact required. Here I focus on articulating the puzzle reflection poses for the rational-ability view in particular.
10. For a more detailed elaboration of this view, see Nelkin (2011, 2016). Wolf (1990) does not use the language of "accountability" or understand responsibility in part in terms of demands, but she offers a similar set of satisfaction conditions for what she calls "free and responsible" action.
11. See Brink and Nelkin (2013).
12. See, for example, Hart (1961).
13. See Lewis (1981) and Beebee and Mele (2002).
14. On the view I favor, having the earlier opportunity does not require any actual *exercise* of agency; but it does require awareness of risk in not acting. For further elaboration and defense, see Nelkin and Rickless (2017). For some foundational discussions of tracing, see Fischer and Ravizza (1998), Vargas (2005), and Fischer and Tognazzini (2009).
15. See, for example, Fischer and Ravizza (1998).
16. See also Vargas (2013, 216).
17. See Coates and Swenson (2013) who propose something like this as an extension of Fischer and Ravizza's account, and see Nelkin (2016) for a reply.
18. While Annas (2011) is concerned to explain moral virtue and expertise, rather than responsible agency, it is notable that on her view expert action involves a "trace" of reflection. See n. 7.
19. Pamela Hieronymi (2014) considers the similar suggestion that one must understand that others have rights and interests to be responsible and that this might offer a role for reflection in responsible agency. Notably, she presents a picture in which an agent's quality of will expressed in an action is the fundamental object of responsibility practices. Thus, she is not addressing The Puzzle set out above for rational-ability views. At the same time, her work illustrates the existence of a parallel puzzle for alternative accounts of responsibility, and a more general need to explain both the appeal of reflection as a necessary condition on responsible agency and intuitions about nonreflective responsible agency. Interestingly, this solution to the puzzle is not available to theorists who also take quality of will to be central and yet do not require moral understanding. (See Scanlon 2008; Smith 2015; and Talbert 2012).
20. See Velleman (2000) for a defense of the distinction and the language of "under the guise of the good."
21. Interestingly, Jaworska's purpose in raising the distinction between two interpretations of "acting for a reason" is to show that certain of those who take quality of will to be sufficient for responsibility will have trouble defending their commitment to the thesis that psychopaths are morally responsible in virtue of their expressing a bad quality of will. This is because it is hard to see how one can express a quality of will if one only acts on a reason in the weaker sense, and yet arguably, psychopaths can only act on reasons in this sense. However, the distinction could be used to help *both* the rational-abilities view and quality-of-will views to accommodate a role for reflection in responsibility. Just how to interpret responsiveness to reasons in rational-abilities views of responsibility has been a surprisingly underexplored question, but one that has a number of important implications, including, as I hope to have shown, for a possible role for reflective ability.
22. Could there be some other systematic means of correction at the first-order level? I am not sure how to rule out this possibility, but I also don't know what would count.
23. For development of this idea, see Nelkin and Rickless (2017).
24. Railton (2014, 846) claims that reflection is "of the utmost importance" to epistemology and morality, insofar as it helps us sort through good and bad intuitive responses. This seems in the spirit of the suggestion here, though Railton is focused here on evaluative intuitions rather than on actions or omissions that might be done for reasons, and is not primarily focused on questions of morally responsible agency in this context.
25. Arpaly (2003) presents this case in the context of responsibility in an influential text.
26. The question arises whether God would then lack the ability to reflect, not needing it to act perfectly well. But if it is true that personhood requires the ability to reflect, and God is a person, then it would still be the case that a perfect being has the ability to reflect, though it is not needed to act well. (Thanks to Ann Whittle for this point.)
27. As Davia (2019) has noted, it might be that any planning agents who think ahead and coordinate plans must reflect in a robust way, as well. This also undermines the idea of intermittency to an extent.

28. Something like the thought that reflective ability makes us *better* rational agents is also expressed by Fischer (2017) in a recent reply to Doris, although he does not tie it to a notion of a normative ideal.
29. It is possible that the suggestion regarding a normative ideal is consistent with—and even complementary to—one of the earlier candidates for a reflective-ability condition after all, depending on what it is to be an ideal. If ideals in some way guide practice, then perhaps they can only do so for beings who have the general abilities that would allow them to conform with the ideal.
30. I develop this argument in Nelkin (2016).
31. Another way of trying to explain away the appeal of a reflective-ability requirement is to posit a *counterfactual* ability requirement on responsible agency, such as the following: Had S reflected on the reasons for and against *A*-ing at t, she would have endorsed the reasons for *A*-ing at t that she acted on. (See, e.g., Arpaly and Schroeder [2012] who consider it before rejecting it.) However, there are two reasons to reject this debunking explanation, especially if one wishes to retain a rational-ability view. The first is that it seems that the counterfactual is true, when true, in virtue of something about the agent in question. Perhaps it is that the agent's reasons have a certain quality, or that the agent bears a special relationship to them, such as that they are especially strong or have a wide application across a number of behavioral domains, or that they play some central role in the hierarchical organization of the agent's reasons. But in that case, it seems that what is really doing the work is not the truth of the counterfactual but its truthmaker. Second, in offering a new necessary condition on responsible agency, the rational-ability view faces a parallel puzzle to the original. Why should the possession of an opportunity ability to act for good reasons entail the truth of such a counterfactual?
32. For criticism of the historical condition, see Eshleman (2001), Mele (2000), and McKenna (2000; 2012). For an early reply from Fischer, see his (2006).
33. Others have objected by means of counterexamples, including that philosophical skeptics about responsibility still appear blameworthy for certain actions.
34. Interestingly, Wolf (1992) took it that her version of the rational-ability view, the Reason View, already *built in* something like a reflective endorsement ability, offered by what she calls "The Real Self View" modeled on Frankfurt's hierarchical desires. See also McKenna and Van Schoelandt (2015) for a defense of a similar thesis. This is an interesting idea, but it would then appear to build reflective ability into the view in question, and then face questions about the empirical data—as well as thought experiments such as the Huck Finn case discussed in the text.
35. Many thanks to Helen Beebee, Randy Clarke, Derk Pereboom, Sam Rickless, Manuel Vargas, and Ann Whittle for very helpful comments on previous drafts that greatly improved the paper, and to Matt Braich and Cory Davia for many helpful discussions of related issues. This paper has its origins in a graduate seminar I taught in 2017 on Deliberation and Reasons-Responsiveness, and I am very grateful to that group which includes William Albuquerque, Henry Argetsinger, Claudia Blöser, David Brink, Rosalind Chaplin, Kathleen Connelly, Emma Duncan, Melissa Koenig, JiMin Kwon, Joseph Martinez, Marcus McGahhey, Joseph Stratman, and Shawn Wang.

REFERENCES

Annas, Julia 2011. *Intelligent Virtue*, Oxford: Oxford University Press.

Arpaly, Nomy 2003. *Unprincipled Virtue*, Oxford: Oxford University Press.

Arpaly, Nomy and Tim Schroeder 2012. "Deliberation and Acting for Reasons," *The Philosophical Review* 121: 209–39.

Beebee, Helen and Al Mele 2002. "Humean Compatiblism," *Mind* 111: 201–23.

Brink, David O. and Dana K. Nelkin 2013. "Fairness and the Architecture of Responsibility," in Shoemaker ed. (2013), 284–313.

Brink, David O. 2013. "Situationism, Responsibility, and Fair Opportunity," *Social Philosophy and Policy* 30: 121–49.

Buckareff Andrei, Moya Carlos, and Rosell Sergi, eds. 2015. *Agency, Freedom, and Responsibility*, New York: Palgrave Macmillan.

Clarke, Randolph 2015. "Abilities to Act," *Philosophy Compass* 10/12: 893–904.

Coates, Justin and Philip Swenson 2013. "Reasons-Responsiveness and Degrees of Responsibility," *Philosophical Studies* 165: 629–45.
Davia, Cory 2019. "Reflection Without Regress," *Pacific Philosophical Quarterly* 100, 995–1017.
Doris, John M. 2015. *Talking to Ourselves*, Oxford: Oxford University Press.
Fischer, John Martin 2018. "On John Doris' *Talking to Ourselves*," *Social Theory and Practice* 44: 247–53.
Fischer, John Martin and Mark Ravizza 1998. *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.
Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68: 5–20.
Hart, H.L.A. 1961. "Negligence, *Mens Rea*, and Criminal Responsibility," in Hart (1968).
——. 1968. *Punishment and Responsibility*, Oxford: Clarendon Press.
Jaworska, Agnieszka 2017. "Holding Psychopaths Responsible and the Guise of the Good," in Liao and O'Neil, eds. (2017), 66–78.
King, Matt and Peter Carruthers forthcoming. "Responsibility and Consciousness," in Nelkin and Pereboom, eds. (forthcoming).
Kornblith, Hilary 2012. *On Reflection*, Oxford: Oxford University Press.
Korsgaard, Christine 2009. "The Activity of Reason," *Proceedings and Addresses of the American Philosophical Association* 83: 27–47.
Lewis, David 1981. "Are We Free to Break the Laws?" *Theoria* 47: 112–21.
Levy, Neil 2014. *Consciousness and Moral Responsibility*, Oxford: Oxford University Press.
Liao, S. Matthew and Collin O'Neil, eds. 2017. *Current Controversies in Bioethics*, Routledge.
McKenna, Michael 2004. "Responsibility and Globally Manipulated Agents," *Philosophical Topics* 32: 169–92.
——. 2012. "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism," *Journal of Ethics* 16: 145–74.
McKenna, Michael and Chad Van Schoelandt 2015. "Crossing a Mesh Theory with a Reasons-Responsive Theory," in Buckareff, Moya, and Rosell, eds. (2015, 44–64).
Montero, Barbara 2016. *Thought in Action: Expertise and the Conscious Mind*, Oxford: Oxford University Press.
Nelkin, Dana Kay 2011. *Making Sense of Freedom and Responsibility*, Oxford: Oxford University Press.
——. 2016. "Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness," *Noûs* 50: 356–78.
Nelkin, Dana Kay and Derk Pereboom, eds. forthcoming. *The Handbook on Moral Responsibility*, New York: Oxford University Press.
Nelkin, Dana Kay and Samuel C., Rickless 2017. "Responsibility for Unwitting Omissions: A New Tracing Account," in Nelkin and Rickless, eds. (2017, 106–29).
——. 2017a. *The Ethics and Law of Omissions*, New York: Oxford University Press.
Pereboom, Derk 2014. *Free Will, Skepticism, and Meaning in Life*, Oxford: Oxford University Press.
Railton, Peter 2009. "Practical Competence and Fluent Agency," in Sobel and Wall, eds. (2009, 81–115).
——. 2014. "The Affective Dog and Its Rational Tale: Intuition and Attunement," *Ethics* 124: 813–59.
Scanlon, T.M. 2008. *Moral Dimensions: Permissibility, Meaning, and Blame*, Cambridge: Cambridge University Press.
Schlosser, Markus E. 2013. "Conscious Will, Reason-Responsiveness, and Moral Responsibility," *Journal of Ethics* 17: 205–32.
Schramme, Thomas, ed. 2014. *Being Amoral: Psychopathy and Moral Capacity*, Cambridge, MA: MIT Press.

Shoemaker, David, ed. 2013. *Oxford Studies in Agency and Responsibility*, Oxford: Oxford University Press.

Smith, Angela 2015. "Responsibility as Answerability," *Inquiry* 58: 99–126.

Sobel, David and Wall Steven, eds. 2009. *Reasons for Action*, New York: Cambridge University Press.

Talbert, Matthew 2014. "The Significance of Psychopathic Wrongdoing," in Schramme, ed. (2014, 275–300).

Vargas, Manuel R. 2018. "Reflectivism, Skepticism, and Values," *Social Theory and Practice* 44: 255–66.

——. 2013. *Building Better Beings: A Theory of Moral Responsibility*, Oxford: Oxford University Press.

Velleman, David 1992. "The Guise of the Good," *Noûs* 26: 3–26.

Vihvelin, Kadri 2013. *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*, Oxford: Oxford University Press.

Watson, Gary 1975. "Free Agency," *Journal of Philosophy* 72: 205–20.

——. 1996. "Two Faces of Responsibility," *Philosophical Topics* 24: 227–48.

Wolf, Susan 1990. *Freedom Within Reason*, New York: Oxford University Press.