Journal of the American Philosophical Association (2023) 762–778 © The Author(s), 2022. Published by Cambridge University Press on behalf of the American Philosophical Association
DOI:10.1017/apa.2022.25

Reasons-Responsiveness and the Challenge of Irrelevance

ABSTRACT: *Carolina Sartorio has criticized the reasons-responsiveness theory of freedom for being inconsistent with the actual-sequence view motivated by the Frankfurt-style cases. Specifically, reasons-responsiveness conceived as a modal property does not pertain to the actual sequence of the agent's action and thereby it is irrelevant to the agent's freedom and moral responsibility. Call this the challenge of irrelevance. In this article, I present this challenge in a new way that overcomes certain limitations of Sartorio's argument. I argue that the root of the challenge is that reasons-responsiveness as an unmanifested modal property seems to be nonexplanatory for the agent's action. I show that reasons-responsiveness theorists will confront this challenge even if they do not endorse the actual-sequence view. Finally, I deflate this challenge with David Lewis's model of causal explanation, showing that reasons-responsiveness is explanatory in virtue of providing information about the causal history of the agent's action.*

KEYWORDS: free will, moral responsibility, reasons-responsiveness, actual sequences, Harry Frankfurt

Introduction

The reasons-responsiveness theory is popular among contemporary compatibilists. According to this theory, reasons-responsiveness, or the capacity to respond to good reasons, is the freedom required by moral responsibility. The term *reasons-responsiveness theory* is sometimes reserved for the account defended by John Martin Fischer (1994) and further developed by Fischer and Mark Ravizza (1998). In this article, I use it in a broader sense such that it also covers other accounts that associate human's rational capacity with free and responsible agency (such as those by Haji 1998; Smith 2003; McKenna 2013; Vihvelin 2004, 2013).

Usually, reasons-responsiveness is understood as a modal property that is analyzed with counterfactual conditionals or possible scenarios: very roughly, an agent S performs an action φ in a reasons-responsive way if and only if S would refrain from φ-ing in a range of possible scenarios in which S has a sufficient reason not to φ. To illustrate, suppose that Emma is reasons-responsive with

Special thanks to Jules Holroyd, James Lenman, Helen Beebee, T. Ryan Byerly, Carolina Sartorio, David Heering, William Hornett, Mu Liu, Yajun Chen, Yingfeng Chen, and several anonymous referees for helpful comments on drafts of this article. I am also grateful to the audiences of the philosophy postgraduate seminar at University of Sheffield in 2017, the Explanation conference at Wuhan University in 2019, and the Free Will and Causality conference at Heinrich Heine University Düsseldorf in 2019.

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

respect to her playing a video game. Her reasons-responsiveness could then be analyzed in counterfactuals such that *if* there were a sufficient reason against playing it (say, Emma's best friend Henry calls her to drive him to the hospital), Emma would stop playing it because of that reason. Call this the *modal conception of reasons-responsiveness*.

Carolina Sartorio (2015, 2016) brings to light a tension between the modal conception of reasons-responsiveness and another influential compatibilist idea that is motivated by the Frankfurt-style cases, namely the actual-sequence view. According to this view, what matters to freedom and moral responsibility is the actual sequence leading to the agent's action. Sartorio suggests that if we take the actual-sequence view more seriously, it will turn out to speak against the modal conception of reasons-responsiveness: since an agent's reasons-responsiveness understood as a modal property cannot be reflected in the actual sequence of the agent's action, it cannot ground the agent's freedom and moral responsibility. In other words, the actual-sequence view renders reasons-responsiveness irrelevant to freedom and moral responsibility. I refer to this problem as the *challenge of irrelevance* to the reasons-responsiveness theory. This challenge is troublesome. Drawing upon the Frankfurt-style cases, several prominent defenders of reasons-responsiveness theory integrate their accounts with the actual-sequence view (such as Fischer and Ravizza 1998; Haji 1998; McKenna 2013). This challenge, hence, may imply that their positions are internally inconsistent. Even worse, the scope of the challenge can be expanded to concern every reasons-responsiveness theorist who endorses the modal conception, no matter buying the actual-sequence view or not. However, based on David Lewis's model of causal explanation, reasons-responsiveness can be explanatory for the agent's action in virtue of providing information about the causal history of the agent's action.

1. A Compatibilist Marriage

A primary motivation for the actual-sequence view comes from the cases that were first devised by Harry Frankfurt (1969). Below is a typical Frankfurt-style case.

> Jones is considering whether to cheat on the exam. Unbeknown to Jones, Black, a resourceful neuroscientist, wants to ensure that Jones cheats on the exam. Black has implanted a tiny chip into Jones's brain without Jones's awareness. With this chip, Black can monitor Jones's brain activities and make Jones decide in accordance with Black's will. Black is going to make Jones decide to cheat only if he discovers that Jones shows any inclination not to cheat. Otherwise, Black will let Jones make his own decision. It turns out that Jones decides to cheat for his own reason.

Many compatibilists think that Jones is morally responsible for his action despite lacking the ability to do otherwise. The intuitive judgment elicited by the Frankfurt-style cases can be characterized as follows:

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

> The Frankfurt-style case intuition: the ability to do otherwise is not necessary for freedom and moral responsibility.

The Frankfurt-style case intuition is prima facie recalcitrant because of the entrenched idea that freedom requires the ability to do otherwise. Therefore, those who share the Frankfurt-style case intuition try to explain this intuition with a deeper principle about freedom, namely the actual-sequence view (such as Fischer and Ravizza 1998; Haji 1998; McKenna 2013; Sartorio 2015, 2016).

> Actual-sequence view: whether an action is free depends exclusively on how the actual sequence of the action unfolds.

This is just a rough characterization of the actual-sequence view. As I show below, philosophers have more specific interpretations of this view in their discussions. Nevertheless, this characterization suffices to explain why Jones is responsible for his action despite lacking the ability to do otherwise: the neuroscientist's possible intervention only happens counterfactually—it makes no difference to the actual sequence of Jones's action and thereby is irrelevant to the freedom of his action.

Drawing upon the actual-sequence view, a new compatibilist model is advanced, namely the actual-sequence compatibilism. (Apart from the Frankfurt-style cases, there are other motivations for the actual-sequence compatibilism: see Sartorio 2018c; Heering 2020, 2021; Kaiserman 2021.) In this model, whether the agent is acting freely is *not* an issue of whether the action is *determined*, but an issue of whether the action is produced *through the right actual sequence*. The important question for the actual-sequence compatibilists is what the right actual sequence amounts to. Naturally, it should be a causal history of action that does not involve any control-undermining factors such as manipulation, brainwashing, hypnotism, addiction, phobia, psychological disorders. This list of control-undermining factors can extend infinitely. Instead of adding the potential control-undermining factors in a somewhat ad hoc way, it would be nice if we could give an account of what features the agent has to have—what positive conditions there are for responsibility (which are undermined by the presence of those factors). Here are how the reasons-responsiveness theory comes into play. It helps to delineate the right actual sequence of action. That is, when the agent is reasons-responsive with respect to her action, her action is produced in a way without those freedom-undermining factors.

There is a happy marriage between the actual-sequence view and the reasons-responsiveness theory. (There are also actual-sequence compatibilist accounts that are developed independently of the reasons-responsiveness theory: see Frankfurt 1971; Watson 1975.) The most influential proposal to combine these two compatibilist ideas is Fischer and Ravizza's mechanism-based reasons-responsiveness account (1998). According to their account, an agent satisfies the control condition for moral responsibility when her action is produced by her reasons-responsiveness mechanism. Fischer and Ravizza propose that to adequately ground moral responsibility, the operative mechanism should be moderate reasons-responsive, which consists of two conditions, namely regular

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

reasons-receptivity and weak reasons-reactivity. Specifically, to be regular reasons-receptive, the operative mechanism must enable the agent to recognize multiple reasons such that those reasons constitute an understandable pattern; to be weak reasons-reactive, there must exist at least one counterfactual scenario in which the mechanism enables the agent to react to a sufficient reason thereof. The important point is that the operative mechanism's reasons-responsiveness is understood as a modal property and cashed out in counterfactual scenarios. Sartorio (2015, 2016) argues that this modal conception is the source of the tension between the reasons-responsiveness theory and the actual-sequence view.

2. A Compatibilist Divorce

2.1 Sartorio's Argument for the Challenge of Irrelevance

Sartorio construes the actual-sequence view as a grounding claim about freedom, according to which freedom is exclusively grounded in the factors pertaining to the actual sequence of the agent's action. Because grounding is typically a transitive relation, a factor can be pertinent to the actual sequence in two senses: either that the factor is itself an element of the actual sequence; or that the factor helps to ground the actual sequence. Hence, the actual-sequence view should be properly interpreted as the claim that freedom is grounded in the actual sequence plus the grounds of the actual sequence (Sartorio 2016: 11, 2018a, 2018b, 2021).

Moreover, Sartorio takes the actual sequence as the actual causal history of the agent's action and argues that the grounding claim about freedom implies a supervenience claim about freedom. That is, the agent's freedom supervenes on the relevant factors pertaining to the causal history of her action. In other words, the factors that enhance or reduce an agent's freedom must be elements of the causal history (or something that contributes to grounding the causal history) of the agent's action. (Given that the causes are abundant in a causal history and that not all the causes are relevant to an agent's freedom, Sartorio suggests that the supervenience claim should be qualified. That is, freedom supervenes on the *relevant part* of the causal history rather than the *whole* causal history.) Imagine that Kate and Tom perform similar actions but Kate is free and morally responsible for her action while Tom is not. According to the supervenience claim, then, a difference must be found in the causal histories of their actions—say, Tom is hypnotized to act while Kate acts on her own reasons. Sartorio expresses this idea in a slogan—'no difference in freedom without a difference in the causal sequence' (Sartorio 2016: 29).

Sartorio contends that the modal conception of reasons-responsiveness falls short of the supervenience claim, thereby contradicting the actual-sequence view (construed as the grounding claim). Consider the following hypothetical scenario devised by Sartorio (2016: 119-20).

> Frank and Insensitive Frank: Frank chooses to shoot his enemy Furt for certain reasons (say, a desire for revenge). Frank is reasons-responsive with respect to his action. For example, if Frank learned that Furt is

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

> the father of five children who depend on him to survive, Frank would refrain from killing Furt. Insensitive Frank is in a situation that is similar to Frank's in many respects. Insensitive Frank chooses to shoot his enemy Furt* for similar reasons. However, Insensitive Frank is not reasons-responsive with respect to his action—even if he learned that Furt* has five children waiting for Furt* to feed them, Insensitive Frank would not refrain from killing Furt*.

A reasons-responsiveness theorist will conclude that Frank acts freely and responsibly while Insensitive Frank does not. If he also commits to the actual-sequence view, as many other reasons-responsiveness theorists do, he will need to explain how the two agents differ regarding the causal histories of their actions (or the grounds of the causal histories). This is where the difficulty arises. Sartorio contends that the modal conception of reasons-responsiveness provides no clues to answer this question. Particularly, the modal conception of reasons-responsiveness seems to allow that Frank's and Insensitive Frank's actions are produced through similar causal histories in all relevant respects, say, their actions are motivated by similar reasons and caused by similar mental states. This violates the supervenience claim.

The case of Frank and Insensitive Frank serves as an intuition pump in Sartorio's argument. It elicits the intuition that reasons-responsiveness, conceived as a modal property, can hardly be reflected in the actual causal history of the agent's action. This intuition points to a tension between the reasons-responsiveness theory and the actual-sequence view: if, as committed by the actual-sequence view, *all* factors relevant to freedom and moral responsibility must pertain to the actual causal histories, then reasons-responsiveness will be rendered *irrelevant* to freedom and moral responsibility. This is the challenge of irrelevance. (The claim that reasons-responsiveness is not relevant to freedom and moral responsibility should not be conflated with the claim that reasons-responsiveness is not necessary for freedom and moral responsibility. The former is a stronger claim than the latter. *X* can be relevant to *Y* even if *X* is not necessary for *Y*. For instance, being familiar with Kant is not necessary for being a good philosopher; however, it is relevant to being a good philosopher.)

Confronting this challenge, Sartorio (2015, 2016) abandons the modal conception of reasons-responsiveness and instead advances an actualist conception. On Sartorio's account, the agent is reasons-responsive (or 'reasons-sensitive' in Sartorio's terms) with respect to his action only if his action is caused by a proper combination of reasons and *absences of reasons*. Accordingly, Frank's action is free because despite being caused by reasons, it is also caused by absences of reasons, such as the absence of Furt's children who rely on Furt to survive; while Insensitive Frank is not acting freely because his action is not caused by those counterpart absences of reasons (in the sense that Insensitive Frank would not refrain from killing even if Furt* was the father of five children). By letting absences enter the causal histories, Sartorio manages to take reasons-sensitivity as a property cashed out exclusively in terms of the actual sequences and thereby solve the challenge of irrelevance. (For similar conceptions of reasons-responsiveness, see also Heering 2020; Kaiserman 2021.)

Nevertheless, the metaphysical price for this account seems to be high. If we allow absences to enter the causal histories, then there may be too many absences figuring in the causal histories. My staying at home for the weekend, for instance, is caused by the absence of interesting movies; equally, it is caused by the absence of brutal intruders driving me away. Though the latter kind of absences is regarded as highly irrelevant, there seems to be no tenable way to distinguish between the relevant absences from the irrelevant ones (for this point, see Beebee 2004). Perhaps the account is worth the price if it is the only way to handle the challenge of irrelevance. However, as I will show below, we can deflate the challenge even if we retain the modal conception of reasons-responsiveness. On balance, the actualist conception seems to be undermotivated. Before laying out my proposal, however, I first point out two limitations of Sartorio's argument and present the challenge of irrelevance in a different but more compelling way.

2.2 The Limitations of Sartorio's Argument

The first limitation of Sartorio's argument is that it appeals to the unclarified notion of actual causal history. In consequence, the criteria of being pertinent to the actual causal history are unclear. Recall that a crucial intuition behind her argument is that reasons-responsiveness as a modal property is not pertinent to the causal history (nor to the grounds of the causal history) of the agent's behavior. Those reasons-responsiveness theorists who hold a more inclusive conception of actual causal history may simply deny this intuition. For example, Fischer and Ravizza have made it clear that they do not exclude modal or dispositional properties from being part of the actual sequence on their account (Fischer and Ravizza 1998: 53, see also McKenna 2013: 154–55; Fischer 2015: 121; Haji 1998: 79).

Sartorio's argument may be improved by imposing some restrictions on the notion of actual causal history. A natural suggestion is that an actual causal history should include actual facts and preclude counterfactual facts (facts that are cashed out in counterfactual conditionals). This suggestion is untenable. For one thing, some factors are typically cashed out in counterfactuals while remaining intuitively relevant to the outcomes of causal histories. To name a few, the laws of nature, or the dispositional properties with causal power. It seems to be unpalatable to exclude all these modal factors from being reflected in the causal histories. More importantly, even Sartorio is not willing to preclude all counterfactual facts from being pertinent to the actual sequences. Recall that Sartorio construes the actual-sequence view as a grounding claim about freedom, that freedom is grounded in the relevant parts of the causal history of the agent's action. Because grounding relations are typically transitive, factors pertaining to an actual causal history can be either elements of the causal history or elements of the grounds of the causal history. As a result, Sartorio admits that even counterfactual facts can be pertinent to the actual sequences if they contribute to grounding the actual sequences. As she points out, there will be no surprises if

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

actual causal histories are grounded in some counterfactual facts since causation itself may be analyzed counterfactually (Sartorio 2018a, 2018b, 2021).

Accepting this hierarchy grounding structure seems to make it even more difficult to identify the pertinent factors of the actual causal history. Now the line between actual facts and counterfactual facts becomes blurry. Sartorio may take the facts about being caused by absences as actual facts even though these facts are further cashed out by or grounded in some counterfactual facts. But it seems that Fischer and Ravizza can make similar remarks on their modal conception of reasons-responsiveness, that the operative mechanism's being reasons-responsive is an actual fact, even though this fact is further cashed out by or grounded in some counterfactual facts. The debate about what is actual or what is pertinent to the actual sequence is in danger of collapsing into a verbal dispute. Therefore, Sartorio's argument, relying on the elusive notion of actual causal history, is far from decisive.

Another limitation of Sartorio's argument is the scope of the conclusion. The challenge is presented as an inconsistency between the modal conception of reasons-responsiveness and the actual-sequence view. Though many reasons-responsiveness theorists are convinced by the Frankfurt-style case intuition and endorse the actual-sequence view, some others do not. The argument is impotent for these reasons-responsiveness theorists who do not buy the reasoning of the Frankfurt-style cases (see Smith 2003; Nelkin 2011; Vihvelin 2004, 2013).

3. Reformulating the Challenge

My argument for the challenge of irrelevance runs in Sartorio's spirit. Nevertheless, it does not involve the elusive notions of actual sequence or actual causal history, thereby circumventing the difficulty with Sartorio's argument. I also expand the scope of the challenge of irrelevance and argue that the challenge should also concern those reasons-responsiveness theorists who do not buy the Frankfurt-style case intuition or the actual-sequence view.

3.1 Reasons-Responsiveness as an Unmanifested Modal Property

Even though we cannot conclusively rule out an agent's reasons-responsiveness from being pertinent to the actual causal history of her action, we still have the seeming intuition that the agent's reasons-responsiveness is difficult to fit into the actual causal history of her action. Why? The plausible answer is that the agent's reasons-responsiveness with respect to her action seems not to be *causally explanatory* for the occurrence of her action. Thus, we should focus on the causal explanation of the agent's action rather than the causal history of the agent's action when framing the question of whether reasons-responsiveness is relevant to her freedom and moral responsibility. (Sartorio [2016] does not make a rigid distinction between causation and causal explanation; Pereboom [2018] suggests focusing on causal explanation rather than causation for the former notion triggers fewer metaphysical controversies; in her reply, Sartorio [2018a] accepts this suggestion.)

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

Why does reasons-responsiveness appear to be causally nonexplanatory for the action? The answer should not merely be that reasons-responsiveness is taken as a modal property because even modal properties such as dispositions and abilities can be explanatory. A more plausible answer is that reasons-responsiveness is understood as an *unmanifested* disposition or an *unexercised* ability. Suppose that a fragile glass falls onto the ground but does not get broken. At first sight, the fragility of the glass seems to be explanatorily irrelevant to the fact that the glass is not broken. (This intuition corresponds to the 'strict model of causal explanation' discussed below.) Likewise, if reasons-responsiveness is not only understood as a modal property but also as a modal property that *is not manifested* in the actual history, then it may be explanatorily irrelevant.

One might suggest that reasons-responsiveness is not a *purely* modal property; rather, it includes a counterfactual part plus an actual part. Accordingly, whenever the agent is φ-ing for an actual reason, he is also exercising his reasons-responsiveness. Even if this is true, reasons-responsiveness theorists will probably add that it is the counterfactual part, or more precisely, the unmanifested part of reasons-responsiveness that does the *main* grounding job for freedom and moral responsibility. Recall the case of Frank and Insensitive Frank. Both of the agents are choosing to shoot their enemies for the *actual* reasons of revenge. We may say that both of them are exercising the actual part of reasons-responsiveness. Nevertheless, it is arguably the counterfactual part that accounts for the difference in their moral responsibility. That is, in counterfactual scenarios where there were sufficient reasons to act differently, Frank would respond to those reasons while Insensitive would not.

3.2 An Improved Argument for the Challenge of Irrelevance

If the issue is rooted in the explanatory status of reasons-responsiveness as an unmanifested property, we can establish the challenge independently of the actual-sequence view. The actual-sequence view is meant to explain the intuition elicited by the Frankfurt-style cases—the agent is morally responsible though his ability to do otherwise is robbed of by the neuroscientist. The tension between the modal conception of reasons-responsiveness and the actual-sequence view is fundamentally a tension between the modal conception and the reasoning of the Frankfurt-style cases. Since notions such as 'actual sequence' and 'actual causal history' are elusive, it would be nice if compatibilists could explain the Frankfurt-style case intuition without appealing to these notions.

Fortunately, there is indeed a more straightforward way to explain the Frankfurt-style case intuition: since the setting of the neuroscientist is irrelevant to the actual explanation of the agent's action, it is irrelevant to the agent's freedom or moral responsibility. This is how Frankfurt explains the Frankfurt-style case intuition in his seminal article. Frankfurt writes, '[w]hen a fact is [in the way as the presence of the neuroscientist] irrelevant to the problem of accounting for a person's action it seems quite gratuitous to assign it any weight in the assessment of his moral responsibility' (Frankfurt 1969: 837). This explanation involves less commitment than the actual-sequence view for it only claims what is irrelevant to

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

the agent's freedom and moral responsibility while keeping silent on what is relevant. This idea can be articulated more precisely with the *irrelevance principle*, which was first presented by David Palmer (2014: 3853) and then quoted approvingly by Fischer:

> Irrelevance principle. If a fact is irrelevant to a correct account of the causal explanation of the person's action, then this fact is irrelevant to the issue of the person's moral responsibility. (Fischer 2015: 122)

Though Fischer contends that irrelevance principle involves ambiguous phrases and requires more qualifications, for the sake of simplicity, I will not go into these nuances and grant that the current formulation is tenable to advance the argument. If irrelevance principle is taken as a more plausible explanation for the Frankfurt-style case intuition, then there is an alternative way to demonstrate the tension between the reasons-responsiveness theory and the reasoning of the Frankfurt-style cases. Consider the following three theses:

> Thesis 1. Being reasons-responsive is essential for a person's being morally responsible for her action.

> Thesis 2. A person's reasons-responsiveness, conceived as an unmanifested modal property, is irrelevant to the causal explanation of her action.

> Irrelevance principle. If a fact is irrelevant to a correct account of the causal explanation of the person's action, then this fact is irrelevant to the issue of the person's moral responsibility.

For a reasons-responsiveness theorist who is convinced by the Frankfurt-style case intuition such as Fischer, thesis 1, thesis 2, and the irrelevance principle all sound plausible. First of all, he must accept thesis 1 since it is the core idea of the reasons-responsiveness theory. Also, he has a strong reason to accept irrelevance principle for it is the best explanation for the Frankfurt-style case intuition. Finally, as argued, thesis 2 seems to be intuitive as well. Now we can conclude that this reasons-responsiveness theorist confronts an internal inconsistency: thesis 1 is incompatible with the conjunction of thesis 2 and irrelevance principle.

This argument has two advantages over Sartorio's. First, it does not involve elusive notions such as 'actual sequence' or 'actual causal history'. Second, it redirects our attention to the explanatory status of reasons-responsiveness (as an unmanifested property), which better reveals the nature of the challenge. However, similar to Sartorio's argument, this argument does not directly put into question the modal conception of reasons-responsiveness; rather, it only establishes the tension between the reasoning of the Frankfurt-style cases and the modal conception of reasons-responsiveness. That is to say, this argument has no impact on those reasons-responsiveness theorists who do not buy the

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

Frankfurt-style cases at the very beginning. Indeed, almost every reasons-responsiveness theorist will confront the challenge of irrelevance.

3.3 Expanding the Challenge: The Explanatory Hypothesis

At the heart of the above argument lies the following inference: if the agent's reasons-responsiveness is nonexplanatory for his action, then it will be irrelevant to his freedom and moral responsibility. This inference is vindicated by irrelevance principle, a principle that connects *being nonexplanatory* for the agent's action with *being irrelevant* to the agent's moral responsibility. Nevertheless, the irrelevance principle is not the only way to demonstrate the connection.

Gunnar Björnsson and Karl Persson (2012) propose a model of responsibility judgment, according to which responsibility judgment is a kind of explanatory judgment, which they call the *explanatory hypothesis*. The basic idea is that when we are ascribing moral responsibility to the agent in question, we are making a judgment about whether the agent's motivational structure plays a remarkable role in explaining his action. Specifically, factors increasing or decreasing the explanatory significance of the motivational structure will influence people's attribution of responsibility correspondingly. This hypothesis is promising for it both accommodates the phenomena of ordinary moral practice as well as our intuitive reactions to many important philosophical arguments in the free will literature. In particular, it accounts for why certain excuses affect responsibility attribution: effective excuses serve as independent explanatory factors that make the motivational structure less explanatory or not explanatory. Examples include such factors as *I am out of control*, *I don't know it*, *I was forced*. Besides, this hypothesis explains the appeal of certain skeptical arguments against moral responsibility. For example, according to Galen Strawson's basic argument, no one is ultimately morally responsible for his action because one's action results from *the way one is* (such as one's character, values, preferences); while *the way one is* comes from remote factors such as heredity, childhood experience, and environmental influences, over which one has no control (Strawson 1994). The explanatory hypothesis explains why we find this argument compelling—it shifts our attention from the agent's motivational structure to more distant explanatory factors such as the agent's heredity or his childhood experience. If we endorse the explanatory hypothesis, we then have a new way to raise the challenge of irrelevance. According to this hypothesis, if reasons-responsiveness is not explanatory for the agent's action, then a fortiori it will not have any impact on the explanatory significance of the agent's motivational structure that produces the action. We then reach the same conclusion that the agent's reasons-responsiveness is irrelevant to the agent's free and responsible action.

The challenge of irrelevance hinges on the following inference: if reasons-responsiveness is nonexplanatory for the agent's action, then it is irrelevant to freedom and moral responsibility. This inference can be vindicated in at least two ways. The first is to invoke irrelevance principle motivated by the Frankfurt-style cases; the second is to appeal to Björnsson and Persson's explanatory hypothesis. Particularly with the explanatory hypothesis, the scope of the challenge of

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

irrelevance is expanded: whether reasons-responsiveness theorists endorse the output from the Frankfurt-style cases or not, they still confront this challenge (if they endorse the plausible explanatory hypothesis of moral responsibility).

4. How Reasons-Responsiveness Can Be Explanatory

Although I offer an improved argument for the challenge of irrelevance and expand the scope of the challenge, there is a way to deflate the challenge and save the modal conception of reasons-responsiveness.

4.1 David Lewis's Model of Causal Explanation

Here is the quick answer to the question of how people get the impression that an agent's being reasons-responsive is not explanatory for the agent's action. That is, reasons-responsiveness is usually taken as an unexercised ability or an unmanifested disposition of the agent. However, this answer is incomplete. The impression is also rooted in a specific model of causal explanation, which I call the *strict model of causal explanation*.

> The strict model of causal explanation: to explain an event E is to cite the proper elements of the causal history that have causal influences on E.

I take this model to be neutral to the metaphysical issues of causation. For example, it is neutral to the debate about the ontological categories of the relata of causal relations (such as facts, events, properties). Besides, it is neutral to the nature of causal influences (such as regularities, counterfactual dependence, conserved quantity). On this model, an unmanifested modal property is not explanatory because it neither has a causal influence on the outcome of the causal history nor corresponds to any factors that figure in the causal history. However, this model is too demanding. As to be shown shortly, we often explain the occurrence of events by citing causally inert factors. Consider a more permissive model of causal explanation introduced by David Lewis:

> Lewis's model of causal explanation: to explain an event E is to provide some information about its causal history. (Lewis 1986: 217)

Lewis's model covers certain cases where factors that are causally inert (on specific metaphysical accounts of causation) are cited in explanans. One application is in demonstrating the explanatory relevance of high-level properties. According to Frank Jackson and Philip Pettit (1990), high-level properties such as multi-realizable dispositions can be cited to explain events though it is the underlying low-level properties that are actually 'causally efficacious'. For example, the fragility of the glass (as a high-level property) can be cited to explain the breakage of the glass. Yet, it is the low-level properties that realize the cited high-level properties—say, the molecule structures that underpin the glass's fragility—that actually do the causal work. If the fragility of the glass is causally

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

inert, how can it be causally relevant? To answer this question, Jackson and Pettit propose the account of program explanation: a high-level property is explanatory if it ensures or 'programs for' a low-level property that is causally efficacious. The core idea of the account, as admitted by Jackson and Pettit, can be rephrased with the terms of Lewis's model. For example, the glass's fragility explains the glass's breakage because the fragility indicates a set of possible causally efficacious molecule structures; among these possible molecule structures, there is one that is actualized and figures in the causal history of the glass's breakage.

Another application of Lewis's model is in accounting for the explanatory relevance of absences. Insisting on the metaphysical view that only events enter causal relations, Helen Beebee (2004) contends that absences do not cause things. Nevertheless, she holds that absences are explanatory and suggests that the common-sense talk of absence by causation should be paraphrased as talk of causal explanation. For example, 'Flora's failure to water the orchids caused their death' (which is a claim of absence by causation) should be interpreted as 'the orchids died because Flora failed to water the orchids' (which is a claim of causal explanation). Beebee faces a similar problem as with Jackson and Pettit: if absences do not cause things, how can they be explanatorily relevant? Inspired by Lewis's model of causal explanation, Beebee proposes that, though absences do not directly provide information about the actual causal processes, they do provide *modal* information related to the causal histories. Absences tell us how the causal process would unfold in the closest possible worlds where the 'actual absent events occurred'. Flora's failure to water the orchids, for example, points to those nearby possible words where Flora remembered to water the orchids and enabled the orchids to continue to flourish. (For an extension of Beebee's account to accommodate the cases of absences as explananda, see Tang 2015.)

I do not mean to get into the metaphysical debates of whether high-level properties or absences have causal influences. The point is that even if those factors are causally inert, they are still explanatorily relevant on Lewis's model. Lewis's model offers hope to those who believe that reasons-responsiveness (as an unmanifested modal property) is explanatorily relevant.

4.2 Reasons-Responsiveness and Causal Information

If we want to apply Lewis's model to account for the explanatory relevance of reasons-responsiveness, we need to answer a crucial question: How does an agent's reasons-responsiveness provides information about the causal history of the agent's action?

One immediate suggestion is to invoke Jackson and Pettit's account of program explanation. Perhaps an agent's reasons-responsiveness ensures certain low-level properties that are causally efficacious. Note that the account of program explanation works well when the explanans involve the presence of a manifested dispositional property (such as the fragility of the glass explains its breakage). However, since reasons-responsiveness is now understood as an unmanifested modal property, it is difficult to see how it corresponds to any low-level properties that figure in the actual causal process.

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

Another suggestion is to borrow the idea from Beebee's account of absence explanation. That is, reasons-responsiveness is explanatory in virtue of providing modal information. Specifically, even though reasons-responsiveness does not directly point to the occurrences in the causal history of the agent's action, it tells us something about the closest possible worlds where the agent acted differently in light of a sufficient reason. However, this suggestion fails to achieve an important dialectical aim—to convince the actual-sequence compatibilists that the modal conception of reasons-responsiveness is in line with the actual-sequence view. If reasons-responsiveness merely tells us something about the closest possible worlds, the actual-sequence compatibilists will probably not regard it as a significant explanatory factor such that it makes a difference to our judgment regarding moral responsibility. More importantly, they will think this is a too dangerous move to take, at least for those who endorse the reasoning of the Frankfurt-style cases. The reasoning hinges on the assumption that Black, as a *counterfactual* intervener, is not explanatorily relevant to the agent's action. If they accept that reasons-responsiveness is explanatory in virtue of providing modal information, they may also need to accept that the counterfactual setting of Black is explanatory, which undermines the reasoning of the Frankfurt-style cases.

A satisfactory proposal to show the explanatory relevance of reasons-responsiveness needs to meet two desiderata: (1) the proposal does not equally render explanatory the counterfactual intervener in a Frankfurt-style case; (2) the proposal not only accounts for why reasons-responsiveness is an explanatory factor, but also why it is an explanatory factor that is significant enough to make a difference to our judgments regarding moral responsibility. To provide a proposal that meets these two desiderata, two assumptions must be put on the table. Below is the first one.

> Assumption 1: A person's reasons-responsiveness is based in part on his motivational mental states (that is, beliefs, desires) being sensitive to reasons and evidence.

For illustration, recall the case of Frank and Insensitive Frank. Consider Frank's action first. Typically, Frank's action is explained by his motivational mental states such as beliefs, desires, and intentions. (I presume the causal theory of action explanation that is defended by Davidson [1963] and many others.) The difficult question is how Frank's reasons-responsiveness can be explanatory if his mental states have already done the explanatory job. The key to answering this question is to avoid taking the agent's reasons-responsiveness as a faculty *over and above* his motivational mental states. Consider the process of Frank's exercising his reasons-responsiveness. Imagine Frank's action of shooting Furt is caused by, among other things, his belief B <Furt should be killed>. In a counterfactual scenario, just before implementing his plan, Frank somehow learns that Furt has five children relying on him to survive. This makes Frank hesitate, reflect on B, and eventually abandon B. Frank's deliberational process hinges on the fact that his belief B is sensitive to reasons and evidence. Likewise, Insensitive Frank's not being reasons-responsive must have something to do with his motivational mental

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

states. Imagine Insensitive Frank's action is caused by, among other things, his belief B* <Furt* should be killed>. We can reasonably presume that Insensitive Frank is not reasons-responsive in killing Furt* because he is stubbornly holding B*. Even if Insensitive Frank was informed that Furt* is the father of five children waiting to be fed, his belief B* would not be revised in accordance with this new information. Note that Insensitive Frank's not being reasons-responsive may be due to other types of malfunctioning mental states, say, an irresistible desire to kill Furt*, whereas a counterpart desire figuring in the causal history of Frank's action is sensitive to reasons and evidence. As Pettit and Smith (1996) point out, as with beliefs, desires should be constrained by normative considerations. (For more detailed elaborations on the connection between an agent's rational capacity and her motivational mental states, see Pettit and Smith 1996; Haji 1998: 75–79; Smith 2003.)

Registering the intimate relationship between the faculty of reasons-responsiveness and the agent's motivational mental states is the first step to establishing the explanatory relevance of reasons-responsiveness. Besides, we need a further assumption.

> Assumption 2: being sensitive to reasons and evidence is a property that contributes to identifying a normal motivational mental state.

Return to the above supposition—Frank and Insensitive Frank's actions are caused by prima facie similar motivational mental states. Frank's action is caused by, among other things, his belief B <Furt should be killed>; Insensitive Frank's action is caused by, among other things, his belief B*<Furt* should be killed>. B and B* are prima facie similar, for both are beliefs with similar propositional content. Nevertheless, as supposed, they are different in one important respect: B is sensitive to reasons and evidence while B* is not. This difference is highlighted through the lens of the functionalist account of belief. On this account, a particular belief is distinguished by its (typically) actual or potential causal relations to the agent's behavior and other mental states. (For a useful outline of the functionalist approach to belief, see Schwitzgebel 2019: §1.4.) In this vein, a belief's being sensitive to reasons and evidence is a functional property that can be cashed out in terms of its potential causal relations to other mental states, i.e., this belief tends to be revised when it is spoken against by certain perceptual states and other beliefs (with justification). From a functionalist perspective, B* and B are located in different causal nexuses and play different causal roles. Hence, Frank and Insensitive Frank's difference in reasons-responsiveness is reflected in the difference in their motivational mental states figuring in the causal histories of their actions. This constitutes an important objection to Sartorio's claim that the modal conception of reasons-responsiveness has no resources to distinguish the causal histories of Frank's and Insensitive Frank's actions.

Being sensitive to reasons and evidence is not a marginal feature of beliefs; rather, it is essential for a belief to function normally. Consider a patient who suffers from Capgras delusion—he cannot get rid of a belief that a friend of his is replaced by an impostor who just looks like that friend. The belief of the patient is not constrained

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

by reasons and evidence so we regard it as being pathological as opposed to being normal. A pathological or malfunctioning motivational mental state typically undermines the agent's freedom and moral responsibility. When an unwilling addict is suffering from an irresistible desire, he is not acting freely and responsibly. This point squares with our judgments of moral responsibility regarding Frank's and Insensitive Frank's actions. In particular, Insensitive Frank would probably be absolved of responsibility for his action because his being not reasons-responsive indicates certain abnormal or pathological mental states in the causal history of his action.

Thus, there are two plausible assumptions regarding reasons-responsiveness. The first is that a person's reasons-responsiveness is based in part on his motivational mental states being sensitive to reasons and evidence. The second is that being sensitive to reasons and evidence is a property that contributes to identifying a normal motivational mental state. With these two assumptions, we can account for the explanatory relevance of reasons-responsiveness with Lewis's model of causal explanation. The fact that an agent is reasons-responsive in performing a certain action provides the information that the action is produced by normal (as opposed to abnormal) motivational mental states which typically figure in action explanation.

This proposal meets the two desiderata set above. First, the causal information offered by reasons-responsiveness is not purely modal—it points to actual occurrences in the causal history of action, namely the agent's motivational mental states (or the corresponding physical realizers). Thus, this proposal will not equally render explanatory the counterfactual intervener in a Frankfurt-style case. Second, on this proposal, reasons-responsiveness is not only an explanatory factor but also an explanatory factor that is significant enough to make a difference to our judgments regarding an agent's moral responsibility, for it indicates the presence or absence of the abnormal or pathological motivational mental states in the causal history of the agent's action that typically undermines moral responsibility.

Conclusion

Sartorio has raised a challenge to the reasons-responsiveness theory that I call the *challenge of irrelevance*. Specifically, the actual-sequence view motivated by the Frankfurt-style cases appears to render reasons-responsiveness irrelevant to freedom and moral responsibility. The tension between the reasons-responsiveness theory and the output from the Frankfurt-style cases is rooted in the impression that reasons-responsiveness as an unmanifested modal property seems not to be explanatory for the agent's action. The connection between explanation and moral responsibility can be established through the explanatory hypothesis. This expands the scope of the challenge: almost every reasons-responsiveness theorist, regardless of buying the reasoning of the Frankfurt-style cases or not, should be concerned by this challenge.

To save the modal conception of reasons-responsiveness, reasons-responsiveness can be explanatory with Lewis's model of causal explanation. Specifically, an agent's

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

reasons-responsiveness with respect to a certain action indicates that the action is produced by normal motivational mental states. Once we recognize the close relationship between an agent's reasons-responsiveness and the functioning of his motivational mental states, we will not worry about the explanatory status of reasons-responsiveness.

JINGBO HU
FUDAN UNIVERSITY
jingbo.hu07@outlook.com

References

Beebee, Helen. (2004) 'Causing and Nothingness'. In John Collins, Ned Hall, and L. A. Paul (eds.), *Causation and Counterfactuals* (Cambridge, MA: MIT Press), 291–308.

Björnsson, Gunnar, and Karl Persson. (2012) 'The Explanatory Component of Moral Responsibility'. *Noûs*, 46, 326–54.

Davidson, Donald. (1963) "Actions, Reasons, and Causes'. *Journal of Philosophy*, 60, 685–700.

Fischer, John Martin. (1994) *The Metaphysics of Free Will: An Essay on Control*. Oxford: Wiley-Blackwell.

Fischer, John Martin. (2015) 'Responsibility and the Actual Sequence'. In David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*. Vol. 3 (Oxford: Oxford University Press), 121–36.

Fischer, John Martin, and Mark Ravizza. (1998) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Frankfurt, Harry G. (1969) 'Alternate Possibilities and Moral Responsibility'. *Journal of Philosophy,* 66, 829–39.

Frankfurt, Harry G. (1971) 'Freedom of the Will and the Concept of a Person'. *Journal of Philosophy,* 68, 5–20.

Haji, Ishtiyaque. (1998) *Moral Appraisability: Puzzles, Proposals, and Perplexities*. New York: Oxford University Press.

Heering, David. (2020) 'Actual Control: Demodalising Free Will'. PhD diss., University of Leeds. https://etheses.whiterose.ac.uk/28373/.

Heering, David. (2021) 'Actual Sequences, Frankfurt-Cases, and Non-Accidentality'. *Inquiry*. Published ahead of print. https://doi.org/10.1080/0020174X.2021.1904644.

Jackson, Frank, and Philip Pettit. (1990) 'Program Explanation: A General Perspective'. *Analysis,* 50, 107–17.

Kaiserman, Alex. (2021) 'Reasons-Sensitivity and Degrees of Free Will'. *Philosophy and Phenomenological Research,* 103, 687–709.

Lewis, David. (1986) 'Causal Explanation'. In Lewis, *Philosophical Papers*. Vol. 2 (New York: Oxford University Press), 214–40.

McKenna, Michael. (2013) 'Reasons-Responsiveness, Agents, and Mechanisms'. In David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*. Vol. 3 (Oxford: Oxford University Press), 151–83.

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

Palmer, David. (2014) 'Deterministic Frankfurt Cases'. *Synthese*, 191, 3847–64.

Pereboom, Derk. (2018) 'On Carolina Sartorio's Causation and Free Will'. *Philosophical Studies,* 175, 1535–43.

Pettit, Philip, and Michael Smith. (1996) 'Freedom in Belief and Desire'. *Journal of Philosophy,* 93, 429–49.

Sartorio, Carolina. (2015) 'Sensitivity to Reasons and Actual Sequences'. In David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*. Vol. 3 (Oxford: Oxford University Press), 105–20.

Sartorio, Carolina. (2016) *Causation and Free Will*. Oxford: Oxford University Press.

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press

Sartorio, Carolina. (2018a) 'Replies to Critics'. *Teorema: Revista Internacional de Filosofía,* 37, 107–22.

Sartorio, Carolina. (2018b) 'Replies to Critics'. *Philosophical Studies,* 175, 1545–56.

Sartorio, Carolina. (2018c) 'Situations and Responsiveness to Reasons'. *Noûs,* 52, 796–807.

Sartorio, Carolina. (2021) 'The Grounds of Our Freedom'. *Inquiry*. Published ahead of print. https://doi.org/10.1080/0020174X.2021.1904643.

Schwitzgebel, Eric. (2019) 'Belief'. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2019/entries/belief/.

Smith, Michael. (2003) 'Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion'. In Sarah Stroud and Christine Tappolet (eds.), *Weakness of Will and Practical Irrationality* (Oxford: Clarendon Press), 17–38.

Strawson, Galen. (1994) 'The Impossibility of Moral Responsibility'. *Philosophical Studies,* 75, 5–24.

Tang, Zhiheng. (2015) 'Absence Causation and a Liberal Theory of Causal Explanation'. *Australasian Journal of Philosophy,* 93, 688–705.

Vihvelin, Kadri. (2004) 'Free Will Demystified: A Dispositional Account'. *Philosophical Topics,* 32, 427–50.

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

Watson, Gary. (1975) 'Free Agency'. *Journal of Philosophy,* 72, 205–20.

https://doi.org/10.1017/apa.2022.25 Published online by Cambridge University Press