DOI: https://doi.org/10.22201/iifs.18704905e.2026.1767

CAUSALISM: A FRAMEWORK FOR MORAL RESPONSIBILITY

Carolina Sartorio
Rutgers University
USA
carolina.sartorio@rutgers.edu
https://orcid.org/0009-0009-3797-6000

SUMMARY: This essay is based on the Gaos lectures given at UNAM in March 2025. The general topic is the metaphysical underpinnings of moral responsibility, both in its basic and non-basic forms. It motivates *causalism* as a general framework from which to understand both forms of responsibility.

It consists of two parts. The focus of the first part is basic responsibility—in particular, the metaphysical condition for that form of responsibility: free action. It motivates a causalist, compatibilist view of free action as an extension of the mainstream causalist conception of action. The focus of the second part is non-basic responsibility—responsibility for outcomes in the world. It motivates a view of the conditions under which responsibility for outcomes results from responsibility for actions, one that works as an extension of the causalist view of free action.

Some highlights of the paper are the following. It provides a comprehensive, unified account of the phenomena. It highlights the role played by key metaphysical concepts (like causation, grounding, and powers) in a theory of moral responsibility. Finally, it discusses subtle methodological issues that arise when dealing with a mix of moral and metaphysical judgments.

KEYWORDS: free will, action, agency, causal control, compatibilism

RESUMEN: Este ensayo se basa en las conferencias Gaos impartidas en la UNAM en marzo de 2025. El tema general son los fundamentos metafísicos de la responsabilidad moral, tanto en su forma básica como no básica. Motiva *el causalismo* como un marco general desde el cual entender ambas formas de responsabilidad.

Consta de dos partes. El enfoque de la primera parte es la responsabilidad básica—en particular, la condición metafísica para esa forma de responsabilidad: la acción libre. Motiva una concepción causalista y compatibilista de la acción libre como una extensión de la teoría causalista dominante de la acción. El enfoque de la segunda parte es la responsabilidad no básica: la responsabilidad por los resultados en el mundo. Motiva una concepción de las condiciones bajo las cuales la responsabilidad por los resultados resulta de la responsabilidad por acciones, una que funciona como una extensión de la teoría causalista de la acción libre.

Algunos de los puntos destacados del artículo son los siguientes. Ofrece una teoría exhaustiva y unificada de la responsabilidad moral. Destaca el papel que desempeñan conceptos metafísicos claves (como la causalidad, la fundamentación y los poderes) en una teoría de la responsabilidad moral. Finalmente, aborda cuestiones metodológicas sutiles que surgen al tratar con una mezcla de juicios morales y metafísicos.

PALABRAS CLAVE: libre albedrío, acción, agencia, control causal, compatibilismo

The topic of this essay is the metaphysical underpinnings of moral responsibility, both in its basic and non-basic forms. It motivates *causalism* as a general framework from which to understand both forms of responsibility. It consists of two parts. The first part is on basic responsibility—in particular, on the metaphysical condition for that form of responsibility: free action (what it is to act from a "free will"). It encourages a view of free action that works as an extension of the mainstream causalist conception of action. This part follows the ideas developed in my recent book, *Causalism* (Sartorio 2023). The second part is on non-basic responsibility—responsibility for outcomes in the world. It motivates a view that works as an extension of the causalist view of free action, and one according to which responsibility for outcomes is inherited from responsibility for actions. This part builds on ideas originally presented in my earlier book, *Causation and Free Will* (Sartorio 2016, chapter 2 especially), as well as those developed in more recent papers (Sartorio (forthcoming a) and (forthcoming b)). The two parts work together as an integrated account of basic and non-basic responsibility.

Although causalism itself is neutral with respect to the classical free will problem, it is particularly well suited for those with compatibilist inclinations (those who believe that it is possible to act freely in a deterministic world). This is so for two reasons. The first is that causalism understands what it is to act freely in terms of the existence of certain causes of our actions (which could be deterministic) rather than in terms of the non-existence of certain causes. The second is that causalism does not understand what it is to act freely in terms of the existence of alternative possibilities. This is an additional plus for the compatibilist given the natural inclination to think that determinism is incompatible with alternative possibilities (and theoretical arguments to that effect—see, notably, van Inwagen 1983). In this respect, causalism sides with the "semi-compatibilism" of Fischer and Ravizza 1998, as well as with other similar views such as McKenna 2013.[1]

Causalism is also particularly well suited for those with naturalistic inclinations. This is because, given how I'm understanding the causalist framework, it only appeals to natural causal processes and

[1] The version of causalism I embrace also understands what it is to act freely in terms of reasons-responsiveness, like these two other views. In other respects, though, there are important differences. As we will see, in virtue of some of these differences, these other views don't count as causalist according to my understanding of causalism.

 DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767

to the ordinary notion of causation between events. It does not appeal to more controversial metaphysical notions such as irreducible powers of "agent-causation" (as in Chisholm 1964 or O'Connor 2000; see Griffith 2017 for discussion) or irreducible ends and teleological explanations (Wilson 1989; Sehon 2016).[2]

In the process of sketching the causalist framework, the essay explores the role that central metaphysical concepts—concepts like grounding, causation, and certain kinds of naturalistic powers—can play in a theory of moral responsibility. It also discusses subtle methodological issues that arise when dealing with a mix of moral and metaphysical judgments.

Given the nature of the discussion, the work is focused on the big picture rather than on the details, by laying out a basic framework that could potentially be filled in in a variety of ways. It also suggests possible avenues for future research on various issues that are left underexplored.

1. *Free Action*

Causalism about free action is an extension of causalism about *action*—the leading view among contemporary philosophers, which was originally popularized by Davidson (1963). Roughly, the causalist view of action is this:

> CA: Actions are behaviors that have the right kinds of causes.

By extension, causalism about free action can be characterized by the following thesis:

> CFA: Free actions are actions that have the right kinds of causes.

In other words, in order for a behavior to be a free action, it must meet two sets of requirements: the requirements that make it an action (and that distinguish actions from behaviors that aren't actions),

[2] Kelley (forthcoming) argues that there is a broader notion of action that includes the behaviors by (possible) beings equipped with powers of irreducible agent-causation or goal-directedness. She suggests the same is true of free action. I think it's plausible to think that agents equipped with such powers could act, and could act freely, by using those powers. However, if (as in my case) the classical free will problem is part of what's driving our investigation, the project we're interested in is one about beings, like us, who arguably only have ordinary powers (for example, we're interested in the question of whether we can act freely in a world where everything we do is determined by factors beyond our control).

DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767 *Crítica*

and the requirements that distinguish free actions from actions that aren't free.

I'll start by summarizing some of the main reasons for embracing causalism about action, which in my opinion make it deserving of the popularity it enjoys nowadays. Then I'll explain how causalism about free action can be motivated in a similar way. Not only can it be motivated in a similar way, but it can be motivated in that way as an *extension* of the causalist view of action—and, I will argue, as its most natural extension.

Davidson, and other causalists after him, noted the following. There is a difference between our actions and the movements of our bodies that are not actions but mere behaviors: intuitively, mere movements are things that happen to us, as passive recipients, but actions are things that we actively do, as agents of such actions. For example, if someone bumps into me and makes me trip, my bodily movement when I trip is not an action but a mere behavior in that it's something that happens to me and not something I do. A natural way to explain what this difference amounts to is in terms of the *causes* of the behaviors. In particular, it's natural to suggest that things that we do (our actions) are behaviors that are caused by internal processes involving mental states of certain kinds (our desires, beliefs, intentions, etc.) whereas things that happen to us (mere behaviors) are behaviors that have causes of other kinds (including external causes such as a collision with another person). This motivates the causalist thought that actions are behaviors that have *the right kinds of causes*.[3]

This is a strong motivation for causalism about action. For it's hard to explain the difference between mere behaviors and actions without appealing to causation. Imagine that one tried to say, for example, that a behavior is an action simply when it matches the agent's intention, without making reference to a causal connection between the intention and the behavior. The problem is that the existence of such a match isn't enough to guarantee that the behavior is an action. I could have planned to move in exactly the way I ended up moving, but, if somebody made me trip before I implemented that intention, my behavior still wouldn't be an action. Importantly, the

[3] There is another sense in which causes must be of the right kind for a behavior to be an action: namely, the *causal chain* linking the relevant mental states to the behavior in question must also be of the proper kind. This is the problem of deviant causal chains, also originally discussed by Davidson, which I'll have to set aside in this article.

 DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767

intention must have also caused the behavior (see, e.g., Mele 2017, chapter 3).

Davidson, and other causalists after him, also noted the following. As rational beings, we typically act for reasons. However, there are times when we have more than one reason to act in one same way. For example, while I really like the taste of artichokes, I also know that they are very nutritious. I have, then, at least two different reasons to eat artichokes. So, imagine that on a certain occasion I eat an artichoke and that the resulting behavior is a genuine action, something that I did for a reason, and not a mere behavior. But which reason is that, if I had more than one? Again, a natural answer appeals to causation: the reason for which I ate the artichoke seems to be, quite simply, the reason that *actually caused* my eating it. This may have been one reason or the other, depending on the case. It might not be clear, even to me, what that reason was. Still, intuitively, the reason for which I acted, and in virtue of which my behavior is not a mere behavior but an action done for reasons, is the reason that actually caused my behavior.

As we have seen, causalism about action can be motivated by reflecting on a couple of illuminating distinctions: on the one hand, the distinction between actions and mere behaviors, and, on the other hand, the distinction between the reasons agents have (in the abstract) for acting and the reasons for which they act. Both distinctions serve to encourage the idea that an action is a behavior that has the right kinds of causes.

I will now suggest that causalism about free action can be motivated in a similar way. This time, it's by appeal to the insights developed by Frankfurt in his classical paper on responsibility and alternative possibilities (Frankfurt 1969).

Here it is important to note that Frankfurt himself didn't embrace causalism about action (in fact, he was one of its critics; see Frankfurt 1978). Still, as we shall see, Frankfurt's approach to free will fits naturally with the causalist framework and its motivations. Consequently, we won't concern ourselves with whether Frankfurt himself would have appreciated the extension of causalism from action to free action (he probably wouldn't have), and we'll focus instead on the plausibility of the ideas themselves and on the internal consistency of the view.

Frankfurt argued against what was, at the time, the most entrenched position on free will: the idea that acting freely requires having alternative possibilities of action, or the ability to do otherwise. While the appeal of this position is undeniable (it's very natural

DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767 *Crítica*

to think of free will in terms of having options or alternatives), Frankfurt argued that this is the result of an illusion. The illusion arises because, when we lack alternatives, that in virtue of which we lack alternatives is typically also what leads us to act. And this results in our not acting freely, for then we don't act for reasons *of our own*. However, Frankfurt argued that the reason why we don't act freely is not, even in such cases, the lack of alternatives, but simply the fact that we don't act for reasons of our own.

To illustrate Frankfurt's reasoning with an example, imagine that A wants B to perform a certain action X, and, to that effect, threatens B with a harsh punishment if B doesn't do X. Imagine that B then does X as a result of A's threat. In that case, A's threat plays a dual role: on the one hand, it is what results in B lacking alternative possibilities, but, on the other hand, it is also what makes B do X. According to Frankfurt, this results in our associating the lack of alternatives with not acting freely. But, again, what explains why B didn't act freely is not the lack of alternatives but the fact that B didn't act for reasons of his own.

One way to see this, Frankfurt suggests, is to think of different situations (which will inevitably have to be quite artificial and unusual) in which the factors that lead us to act come apart from the factors that make our actions inevitable. For example, let's imagine that B had very strong independent reasons for doing X, and it is those reasons (not A's threat) that led him to do X. In that case, B appears to act freely despite the lack of alternatives. Intuitively, this is so because B does what he wanted to do, based on his own reasons for doing X. This suggests that what determines whether agents act freely is, not whether they had alternatives, but the factors that in fact led them to act. And note that the most natural way to interpret this is in causal terms, namely, as the claim that, when an action is free, what makes it free is the fact that it had the right kinds of causes.

To bolster his argument, Frankfurt devised cases (now known as "Frankfurt cases") in which the agent who acts freely, B, is not even aware of the fact that he lacks alternatives because he is unaware of A's existence. In that case, the causes of his behavior cannot be anything other than his own reasons for doing X. In one of the original examples given by Frankfurt, A is a neuroscientist who secretly has access to B's mental processes, can predict his behaviors at every moment, and can intervene, if necessary, to force B to make the decision to do X. In the actual case, A doesn't need to

 DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767

intervene because he foresees that B will act for his own reasons. Again, in these circumstances B seems to act freely despite lacking alternatives—intuitively, because his behavior has the right kinds of causes.

As we can see, there is an important connection between Frankfurt's argument about the nature of free action and Davidson's argument about the nature of action (a connection that seems to have been overlooked in the literature—possibly because, as I mentioned, Frankfurt himself rejected causalism about action). The connection concerns the importance of the *actual causes* of behaviors. On the one hand, there is Davidson's idea that what makes a behavior an action, or an action done for reasons, are the factors that actually resulted in the behavior, as opposed to other factors that could have issued in the behavior (such as external factors or reasons that didn't actually move the agent to act). On the other hand, there is Frankfurt's idea that what determines whether an action is free, in the sense relevant for moral responsibility, are the factors that actually moved the agent to act, as opposed to factors that could have, under other circumstances, moved him to act (factors such as a threat by another agent, or the presence of the neuroscientist). Both ideas highlight the importance of the actual causes of behaviors. They suggest that what determines whether a behavior has the relevant status (as an action or free action) are the actual causes of the behavior. That is: not the causes that could, in other hypothetical circumstances, have given rise to the behavior, but the causes that *in fact* gave rise to the behavior.

Not only can the two causalist views be motivated in a similar way, but the view about free action can in fact be seen as an extension, and as the *most natural* extension, of the view about action in general. This is also a point that has been overlooked in contemporary discussions on these topics. In this case, this is probably because, although Frankfurt-inspired views have gained much popularity in recent years, they are still considerably less dominant than causalism about action in general. In fact, Frankfurt's argument is typically seen as a reaction to the "classical" and perhaps most intuitive view about free will: the alternative possibilities view. The alternative possibilities view still seems to be regarded as the default position (the one we should accept unless given good reason to think otherwise). This is probably why much effort has been spent trying to figure out if Frankfurt cases are dialectically successful in undermining the alternative possibilities paradigm. The thought seems to be that, unless

one can prove that the classical paradigm fails, there is no reason to look elsewhere.

But I believe that this, too, is a mistake. Just as it is a mistake (if Frankfurt is right) to think that acting freely requires alternative possibilities, it is also a mistake to think that the conception in terms of alternative possibilities deserves to be seen as the default. The position that deserves that distinction is, in fact, the causalist position. The reasoning behind this is simple (and it itself has to do with simplicity considerations). It is this: first, note that pretty much everybody accepts that the actual causes of our acts matter to whether we act freely. If so, this means that we're already committed to the relevance of actual causes. But then any position that holds that *other* things matter too, such as alternative possibilities, is a more complex position about the nature of free action. The view in terms of just actual causes is simpler and thus, arguably, it should be our starting point.

As an illustration, consider the most prominent contemporary advocates of the classical position, such as van Inwagen (1983) and Kane (1996). They both accept the relevance of actual causes to free will (and they do so for good reasons). Both understand free will as a *dual* kind of control, which involves control over both the actual action and at least one alternative possibility. This is important to guarantee, not just that agents *would* have been in control if they had done otherwise (which is required to have robust alternative possibilities), but also, more fundamentally, that they *are* in control when they act in the actual world. On the face of it, this is a more complex model of free will than one in terms of a single form of control.[4] Thus, my suggestion is that we should see if we can explain everything that we need to explain in terms of a single form of control and we should only look elsewhere if it turns out that we can't.

Admittedly, it's tricky to say exactly what "simplicity" amounts to, and sometimes it's hard to make comparisons between specific theories (as when a theory seems to be simpler than others in one respect but more complex in other ways). As we will see next, the development of the causalist view can itself give rise to some theoretical complexity. However, I hope it will become clear that this complexity is not added theoretical baggage but is already plausibly built into

[4] Other libertarian views that require alternative possibilities as an addition, not as a replacement of actual causal control, include those of Franklin 2018 and Law 2022. Someone who does seem to think that actual causal control is not required is Palmer 2021. This is clearly the exception, not the rule.

 DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767

an understanding of ordinary human agency—in particular, of what explains why we act when we act in ordinary circumstances. If I'm right about this, this reinforces the claim that causalism should be our starting point.

How *could* freedom arise from just actual causal control? The key here is to distinguish between less robust and more robust forms of such control. As we have seen, causalism conceives of free action as a special kind of action: action that is caused in the right way. On this picture, the type of causal control required by free actions subsumes and exceeds the type of causal control required by actions that are not free. For example, acting freely requires a more robust form of causal control by the agent than is required to act compulsively. But both, since they are actions, require a more robust form of causal control than mere behaviors.

This makes sense if we think about it in terms of sensitivity to *reasons*, in particular. When agents act compulsively, the causal history of their behavior includes things like the agents' desires, beliefs, and intentions. But it's still an impoverished causal history, especially when compared to the causal histories of free actions, which reflect a sensitivity to a wider range of rational considerations.

To illustrate, imagine someone—call him "Artie"—whose love for artichokes is not healthy: he has a compulsion to eat them, and when he eats them he doesn't eat them freely. Artie's compulsion kicks in when he attends a party where artichokes are served. When Artie eats an artichoke at the party, the causal history of his act contains his actual reasons for eating it (including his desire to eat an artichoke), but that is almost the full extent of his sensitivity to reasons.

In contrast, imagine that Artie has an "ordinary" twin brother who also likes artichokes but who is not compelled to eat them. When Artie's twin eats an artichoke at the party, the causal history of his action reflects a more robust sensitivity to rational considerations. In addition to his responding to his actual reasons for eating the artichoke (such as his desire to eat something tasty, or his desire to eat something nutritious, etc.), Artie's twin is also sensitive to other rational considerations that explain why he eats the artichoke on that occasion. For example, he might eat it partly because he has no reason to believe that it'll be hard on his stomach—after all, he probably doesn't have the disposition to eat something, no matter how tasty or nutritious, if he thinks it'll be hard on his stomach. Similarly, he probably eats it partly because he has no reason to believe that the food has been poisoned by the host, or

DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767 *Crítica*

that there's a financial reward for not eating it, etc. These are all prudential considerations to which he might be sensitive when he eats it. But there are probably moral considerations too: he might eat it partly because he has no reason to believe that others want to eat that very same artichoke, or that eating it will cause others harm, etc. Facts of these kinds reflect a sensitivity to rational and moral considerations that far exceed that of Artie's, who is not sensitive to reasons in the same way due to his compulsion. Note that what these considerations have in common is that they are all absences of reasons to refrain from eating the artichoke. (In what follows I call these reasons "counterincentives", for short.) Unlike Artie, Artie's twin is responding to a robust pattern of absences of counterincentives.

This needn't be at the conscious level, of course. A lot of things affect us without our being conscious of their influence on us, and in ways that matter to our rational and moral agency. Consider actions that we perform habitually, on a regular basis. While driving to work on an ordinary day, I may turn right, as I always do, at a certain intersection. In doing that, I'm sensitive to certain aspects of my surroundings in virtue of which my behavior is rational and/or moral—such as the absence of obstacles blocking the road, the absence of pedestrians crossing the road, etc. In many cases, I'm not conscious of the fact that I'm responding to those things; still, I am sensitive to those things, and in a way that contributes to my rational and moral agency. Similarly, when Artie's twin takes the artichoke from the platter at the party and eats it, he's responding to the absence of various counterincentives (prudential and moral reasons to refrain from acting in that way), even if he's not consciously registering any of this.

Of course, he *could* become conscious of it, if he were to somehow start reflecting on what he's doing, or if he were to start deliberating about his reasons to eat or to not eat the artichoke. But typically he won't be conscious in this way, as there is no need to. Similarly, when I turn right at the familiar intersection, I could become conscious of what I'm doing, or I might start deliberating about my reasons to turn or to not turn, but typically I won't be, as there is no need to. Clearly, on reflection, the cases where we act freely after consciously weighing our reasons are more the exception than the norm. This is good for our mental health, and it doesn't make us any less rational—or free.

I have motivated a particular kind of causalism about free action, one that appeals to causal histories that reflect the agent's reasons-sensitivity. Roughly, this is the view:

 DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767

> RS: When agents act freely, their reasons-sensitivity is manifested in the causal history of their acts, which contains the relevant incentives to act (reasons to act in that way) as well as *the absence of a range of counterincentives* (reasons to refrain from acting in that way).

Of course, none of us are perfectly sensitive to reasons, and we can still act freely; in turn, even some of the most compulsive agents are able to respond to some extreme counterincentives. So, the difference between agents who act freely and agents who don't is not that free agents respond to all considerations of these kinds and non-free agents respond to none of them, but only that free agents respond to enough of them. What counts as enough, though, is not perfectly precise (but neither is the concept of freedom). And it's something that depends on the type of action at issue.[5]

Note that what I'm suggesting here is that free agents are *actually* sensitive to considerations of this kind. This is an important feature that distinguishes the causalist account I'm proposing from other, otherwise similar, reasons-sensitivity accounts of free will that don't presuppose alternative possibilities—such as, most notably, the "semi-compatibilism" of Fischer and Ravizza 1998, but see also McKenna 2013. Those are *counterfactual* accounts of reasons-sensitivity: they explain reasons-sensitivity and thus freedom *directly* in terms of counterfactual facts of certain kinds. (Fischer and Ravizza explain it in terms of the counterfactual behavior of the actual "mechanism" of action, and McKenna explains it in terms of the counterfactual behavior of the agent.) In contrast, the causalist version of the view explains it in terms of actual reasons-sensitivity facts. So, part of this is familiar territory but part of it is not, and the part that is not is in fact what makes it a causalist view. (Those other views are *not* causalist views, as I am understanding causalism.)[6]

[5] An alternative here is to go for a degreed notion. Thus, Kaiserman (2021) uses the reasons-sensitivity framework expressed by a principle like RS, together with his own view on how causal contributions come in degrees, to argue that acting freely comes in degrees. For further discussion of degrees of responsibility within a reasons-sensitivity framework, see Coates and Swenson 2013; Nelkin 2016; Tierney 2019.

[6] As I explain later in this section, causalism is compatible with the existence of counterfactual grounds of the actual reasons-sensitivity facts. But, even if there were such counterfactual grounds, the grounding structure would still go *through* the facts about actual causes or explanations. In contrast, the counterfactual views of Fischer and Ravizza and McKenna attempt to explain freedom directly in terms of counterfactuals.

Another way in which we're stepping out of more familiar territory is that we're making use of some important metaphysical concepts. Most obviously, given that this is a causalist account, causation plays a central role in it. As we have seen, it plays a role in the account of what it is to act, and in the account of what it is to act freely, as the relevant form of causal control is required both to act and to act freely.

Now, here there is a potential complication, and one that concerns both the nature of action and free action. The complication has to do with the question of how to incorporate omissions, and absences more generally, into the account. Arguably, a causalist theory of action should be able to accommodate the agency expressed not just in our actions but also in our omissions (the things we fail to do, especially when we fail to do them intentionally). For example, just like it should be able to accommodate my voting on a certain election, it should also be able to accommodate my failing to vote. But omissions are arguably absences (absences of actions of certain types), and thus when applied to omissions causalism seems to presuppose that absences can be among the causal relata. Similarly, the causalist theory of free action that I offered makes use of sensitivity to absences of certain other kinds (absences of reasons), which also seems to incur a commitment to causation involving absences. But absences are ontologically suspicious creatures. As a result, whether causation involving absences is even possible is a highly controversial issue.[7] So, what if it's not possible?

It's important to see that causalism, or something very close to causalism, can still be true even in that case. Even if absences cannot be causes, and regardless of how exactly we make sense of them from an ontological point of view, they clearly play a role of some kind or other in the explanations of events. For example, a drought is arguably explained partly in terms of the lack of rain, regardless of how we understand such a lack, and regardless of whether it's a genuine cause of the drought. Also, note that, if the explanations in question are not causal, they are still "ordinary" in an important sense: they don't appeal to anything like irreducible ends, in the way teleological explanations do, but only to antecedent states of the world.

Moreover, and perhaps most importantly, they still involve causation at a different juncture. For they are themselves accounted for in terms of counterfactual facts about causation. For example, if

[7] See Bernstein 2015 for an overview of this debate.

 DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767

the lack of rain doesn't cause the drought but helps explain why it occurred, presumably this has something do with facts about hypothetical scenarios where it rains and where this causes the ground to be wet and fertile. In fact, this is how authors who reject the causal powers of absences tend to understand how they can otherwise figure in explanations (see, e.g., Dowe 2001; Beebee 2004). Dowe, in particular, sees a strong resemblance between the powers of absences and causal powers and thus calls them "quasi-causal".

In what follows we'll borrow this term from Dowe. All of this suggests that causalists can rely on a fallback view, which I'll call *quasi-causalism*, which doesn't hinge on the possibility of absence causation. The view is a close cousin of the causalist view:

> QA: Actions are behaviors with the right kinds of (quasi)causes.
>
> QFA: Free actions are actions with the right kinds of (quasi) causes.

Self-proclaimed causalists should (and I suspect would) be happy with quasi-causalism. For quasi-causalism seems to preserve the "essence" of causalism while remaining neutral on a metaphysical debate that causalists should not take a stand on.

Here we see the bearing of a second metaphysical concept: *grounding*. As I'm understanding the project of accounting for our free agency, we aim to discover the metaphysical grounds of the facts about free action. But there are, at least potentially, different levels of metaphysical grounds. Most obviously, there is a first level of direct or immediate grounds. This is a set of *full grounds* for the facts about free agency. But, unless those are metaphysically primitive facts, there is also a level consisting of the grounds of those grounds, and the grounds of the grounds, etc. So, assuming grounding is transitive (or, at least, assuming that these aren't cases that exhibit the intransitivity of grounding), the project of accounting for our free agency can potentially yield more than one level of grounds.

If absence causation is possible, the hierarchy of grounds looks like this:

> Level 0: Free action fact (S acted freely by A-ing).
>
> Level 1: Causal facts (S's A-ing had the right kinds of causes).
>
> Level 2: . . . (The grounds of level 1 facts).
>
> . . .

If absence causation is impossible, the hierarchy is only slightly different starting at level 1:

> Level 1: Quasi-causal facts (S's A-ing had the right kinds of quasi-causes).
>
> Level 2: Causal facts (the counterfactual causal facts that ground the quasi-causal facts).
>
> Level 3: . . . (The grounds of level 2 facts).
>
> . . .

This gives rise to the question: Which levels of grounds is the action theorist interested in, when giving an account of action or free action? I take it the answer is, at least typically: the levels that are higher up in the hierarchy and not at the bottom of it. In particular, I take it that a theory of action, or of free action, should remain neutral (or as neutral as possible) on controversial metaphysical issues such as the nature of causation in general. A causalist about action or free action (just like a causalist about knowledge, or perception, or reference) is not attempting to give a theory of causation but is only making use of the concept of causation to give a theory of action or free action (or knowledge, or perception, or reference). As a result, I take it that if a theory of action or free action manages to remain neutral on a controversial general metaphysical issue such as the causal powers of absences, this is a virtue of the theory. For it means that it doesn't rest on a metaphysically shaky foundation.[8]

Now, at this point someone might worry that a more permissive theory of this kind will end up appealing to more than just *actual* facts. However, wasn't causalism motivated by the thought that only actual explanations matter (to both acting and acting freely)? But this worry is unfounded. For, again, causalism is motivated by this thought only as far as the more direct grounds are concerned; it is not committed to the idea that no other facts appear elsewhere in the hierarchy of grounds.

Still, one might think that there is something perverse about the idea that a causalist view, which is motivated by the thought that only actual explanations matter, could end up appealing to the relevance of *counterfactual* causal processes. Wasn't Frankfurt, in particular,

[8] Although this has always seemed clear to me, I was surprised to encounter quite a bit of resistance to the claims of this paragraph. I argue for this in more detail in Sartorio 2022.

set on denying the relevance of counterfactual facts? But this worry is also unfounded. For Frankfurt's insight is in fact compatible with the relevance of certain counterfactual facts. What it's not compatible with is the relevance of alternative possibilities. But, as it's been commonly noted in the literature, having alternative possibilities, in the relevant sense, requires more than the mere truth of certain counterfactuals or the existence of certain possibilities. Intuitively, it requires a certain kind of access to those possibilities, or the ability/opportunity to make them actual, in virtue of which those possibilities are genuinely up to the agent or within the agent's control (see, e.g., Clarke 2009; Whittle 2010; Cyr 2017; Jaster 2022).

By the way, an independent way to see that causalism is compatible with the relevance of certain counterfactual facts is this. Some causation theorists believe that causation itself is grounded in facts about counterfactual dependence—this is a prominent tradition in the metaphysics of causation that started with Lewis 1973. But, if causation is grounded in counterfactual dependence, the counterfactual facts that ground the facts about causation will themselves be grounds—unless, again, this is an instance where the transitivity of grounding fails. (These facts will appear in either level 2 or 3 in the above hierarchies.) But, surely, the fact that causation is grounded in counterfactual dependence is not a reason to think that causalism is false.[9]

As we have seen, it is not within the purview of action theorists to try and settle general metaphysical disputes about concepts like causation. And this means that causalists are not typically interested in the bottom levels of grounds. However, there could be other levels of grounds that fall within the purview of causalists. In particular, if other concepts related to our agency made an appearance in the hierarchy of grounds, this would seem relevant to causalists.[10]

The theory of free action that I motivated above, articulated in terms of the RS thesis, leaves it open that there could be such concepts. Some dispositional concepts, in particular, seem to be good candidates for playing that kind of role. For one may ask: *Why* are

[9] Similarly, Kelley (forthcoming, section 5) draws attention to the fact that explanation is likely modally grounded. However, her account of action posits a modal condition on action that is separate from an explanatory condition (and all of this is arguably at the first level of grounds). In contrast, I am suggesting that a certain explanatory condition is all that's needed, although it may itself be modally grounded.

[10] I discuss this also in Sartorio (forthcoming c), as part of my response to Metz (forthcoming).

DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767 *Crítica*

free agents sensitive to the absences of certain counterincentives when they act? And a natural answer seems to be that this is because of their dispositional qualities. For example, it might be that, when Artie's twin takes the artichoke from the platter at the party, he is being sensitive to the fact that nobody else wants it. According to the version of causalism that I've proposed, this is captured by a (quasi)causal fact: his action of taking the artichoke has, among its (quasi)causes, the fact that nobody else wants it. But why does that (quasi)causal fact obtain? Intuitively, this is because Artie's twin is disposed to refrain from taking the artichoke when somebody else wants it.

Note that this is a power that he has but that he doesn't get to exercise in the actual situation (where he has no reason to believe that someone else wants the artichoke). So, the fact that he has that unmanifested power seems to ground the fact that his act has a certain kind of (quasi)causal history, and, indirectly, the fact that he acts freely. Assuming, for simplicity, that absences can be causes, the grounding hierarchy would look like this:

> Level 0: Free action fact (Artie's twin acted freely when he took the artichoke).
>
> Level 1: Causal facts (His taking the artichoke had the right kinds of causes).
>
> Level 2: Dispositional facts (He had the power to respond to the relevant reasons).
>
> . . .

Unlike facts about the general nature of causation, these facts do seem to fall within the purview of a theory of agency and free agency. For powers of reasons-sensitivity are certainly among the concepts that action-theorists are interested in.

If a causalist were to embrace this addition, the resulting view would be an "enriched" causalism: a view according to which both actual causes and unmanifested powers are relevant to free will.[11] One could even argue that these unmanifested powers represent an ability

[11] As discussed in Sartorio 2023, chapter 4, there are different ways of understanding the role of powers in a causalism thus enriched. Those different ways arise from different ways of understanding powers themselves (which is another issue that the causalist might want to remain neutral on). It matters, for example, whether one embraces a reductive view or a non-reductive, functionalist view.

 DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767

to do otherwise that free agents have, and that is relevant to their acting freely when they so act.[12] Note, however, that this is a quite *general* ability. As a result, a causalism enriched with unmanifested powers in this way is not a view that advocates the significance of alternative possibilities, given how alternative possibilities are commonly understood. This is for the kinds of reasons mentioned above: having genuine alternative possibilities requires more than having a general ability to do otherwise. Arguably, it requires something like a specific ability to do otherwise, or the opportunity to exercise the general ability in the circumstances the agent is in. Clearly, Artie's twin lacks that more specific ability, given that the relevant reasons to which he is responsive are absent in the actual circumstances.

Still, a causalism enriched with unmanifested powers manages to reconcile two different insights about free action. One is the insight inspired by Frankfurt's reasoning, according to which facts about free action are fully grounded in facts about (quasi)causation. The other is the thought that unexercised powers (in particular, powers to do otherwise) are relevant to acting freely, which is an important motivation for the alternative possibilities view. As a result, causalism enriched with unmanifested powers can capture part of that motivation without incurring a commitment to genuine alternative possibilities. This strikes me as a virtue of the view.

In this first part I motivated causalism about free action by drawing attention to the connection with causalism about action and to some common motivations behind both views. I also explained the role that certain metaphysical concepts—in particular, (quasi)causation, grounding, and powers like dispositions and abilities—can play in a theory of that kind.

2. *Responsibility for Outcomes*

In this second part I'll discuss how causalism can be extended from basic to non-basic responsibility. In what follows I assume that responsibility of the non-basic kind is genuinely possible—in other words, that agents can be responsible, not just for their actions or their choices, but also for the outcomes of those actions and choices. Although this certainly aligns with commonsense, it seems to carry

[12] This might depend on how one thinks about the relation between abilities and dispositions—in particular, whether abilities are reducible to dispositions. For arguments that they are, see Vihvelin 2013, chapter 6, and Fara 2008. For arguments against this idea, see van Inwagen 1983, pp. 10–11; Clarke 2015; Vetter and Jaster 2017; Vetter 2019; Wallace 2023.

DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767 *Crítica*

with it a commitment to a type of moral luck (luck about the consequences of our actions) that some find problematic. I'll have to sidestep this debate here (but see Nelkin 2025 for an overview).

By "non-basic" responsibility I just mean responsibility that is not basic. Those things for which we can be responsible in a non-basic way can still be actions of ours. For example, a drunk driver can be non-basically responsible for the action of hitting a pedestrian by virtue of being basically responsible for the earlier action of choosing to drink and drive. I will refer to all those things that we can be responsible for in a non-basic way as "outcomes", but this is just a terminological choice.[13]

First, let me discuss how *not* to extend causalism to non-basic responsibility. Some think that an account of the conditions for non-basic responsibility should mirror, as much as possible, those for basic responsibility.[14] But this strikes me as wrongheaded. Freedom, in particular, is arguably a component of basic responsibility but not of non-basic responsibility. And this means, in particular, that one shouldn't analyze the non-basic form of responsibility in terms of a freedom condition and an epistemic condition, as is common to analyze the basic form of responsibility.[15] More generally, there is no good reason to think that the conditions for the non-basic form of responsibility should mirror the conditions for the basic form.[16] As an analogy, think about non-basic actions and how they're connected to basic actions, according to causalism. We don't expect them to have the same type of structure; for example, non-basic actions aren't the result of a separate intention (my turning the knob and my opening the door, which I do by turning the knob, are both accounted for in terms of a single intention and the same underlying reasons).

How else should one analyze non-basic responsibility, then? A natural answer is: as responsibility that is inherited from more basic forms of responsibility. This is also similar to how we think about actions themselves: non-basic actions are actions that are "derived from" basic actions (in some way or other). In what follows I explain

[13] Kelley (2024) argues for a view of non-basic actions that is more permissive than usual and suggests that this is likely to have important consequences for how we think about our agency. This might be right. But one of those consequences is not, I think, that we're more likely to be responsible for those things (or that we're more responsible for them) simply because they count among our actions.

[14] See, e.g., Fischer and Ravizza 1998, chapter 4, and Capes 2023, chapters 1–2.

[15] I argue for this in Sartorio (forthcoming a).

[16] I argue for this in Sartorio (Ms).

how this type of analysis might go for non-basic responsibility. Here, too, I'm only interested in the main structure of the view, not in the details.

It will help to start with a first pass and then tweak it where needed. At least for the moment, I'll restrict my attention to the simplest cases: cases where the agent's responsibility for an outcome is derived from an individual behavior for which the agent is basically responsible. Consider, for example, this case:

> *Button*: You know that pressing a certain button would cause harm but you still decide to press it, just because you like pressing buttons.

I'm assuming that your responsibility for the harm in this case is derived from your (basic) responsibility for choosing to press the button. What general principle can be behind this? This seems like a natural first answer:

> (P1) S is responsible for an outcome O by virtue of behavior B when S is basically responsible for B and B resulted in O, in roughly the expected way.

In other words, you're responsible for the harm because you're responsible for choosing to press the button (you did it freely, etc.) and because that choice resulted in the harm, in roughly the way you expected.

The modifier "in roughly the expected way" helps to rule out cases where we presumably don't want to hold the agent responsible for the outcome. These are cases of two types: unexpected consequences (imagine that you were completely unaware that pressing the button could cause harm)[17] and cases of causal deviance (imagine that pressing the button in fact resulted in the harm, which was expected, but not at all in the way that you, or anyone else, could have expected).[18]

The principle appeals to the expression "resulted in". Here, too, a causal interpretation is the most obvious one. But, again, we

[17] What if you were culpably unaware? In that case, you could still be responsible for the harm, but your responsibility for the harm would be derived from your basic responsibility for some other earlier action or omission (in virtue of which you were culpably unaware), not from B itself. For more on this general "tracing" strategy, see, e.g., Smith 1983.

[18] For discussion of how causal deviance affects our responsibility for outcomes, see Feinberg 1970.

DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767 *Crítica*

shouldn't just assume that every consequence that we can be responsible for is a causal consequence. This is another instance where we should at least leave it as an open question whether there could be other kinds of consequences for which we're responsible. At least in principle, these could include, again, consequences of our omissions (or absences that are consequences of our actions), if absences are not among the *causal relata*. It could also include non-causal consequences of other kinds. For example, if the harm you caused by pressing the button includes the fact that you turned a man's wife into a widow, that might be a non-causal consequence of your act (because it's a non-causal consequence of a causal consequence of your act—the man's death), and one for which you could be responsible.[19]

Causalists should probably remain neutral on this issue too, as it concerns the general question about the nature of the *causal relata*. Again, we can interpret causalism broadly enough to accommodate the non-causal consequences of our acts. As with absence quasi-causation, this addition doesn't seem to go against the spirit of the causalist view. Note, in particular, that any such non-causal consequences would arguably still be grounded in more basic causal ones. For example, your turning the man's wife into a widow in the button case is partly grounded in your causing the man's death.

Our first pass needs some tweaking, though. Consider the following case:

> *Plane*: The pilot of a falling plane diverts it from a densely populated area to a sparsely populated area, which foreseeably results in the deaths of some people on the ground (in the sparsely populated area).

Imagine that the pilot is responsible—in fact, praiseworthy—for diverting the plane. This led to the deaths of the people on the ground, in the expected way; however, the pilot isn't *praiseworthy* for those deaths. But note that, if we understand our first pass as applying to all kinds of responsibility including praiseworthiness, the account entails that the pilot is praiseworthy for the deaths. For he is praiseworthy

[19] This means that the traditional way of understanding the connection between responsibility and causation, according to which responsibility entails causation, might need to be relaxed (for a recent defense of the traditional view, see Kaiserman (2024)). As I explain next in the text, relaxing that connection is compatible with at least the spirit of causalism.

for diverting the plane, and that action resulted in those deaths, in roughly the expected way.

This motivates a restriction to specific "valences" of responsibility (as in praiseworthiness versus blameworthiness) and types of outcomes that match those valences. Consider, for example, the following principle about *blameworthiness* specifically:

> (P2) S is blameworthy for (a harm) O by virtue of B when S is blameworthy for B and B caused O, in roughly the expected way.

This principle doesn't entail that the pilot is blameworthy for the deaths in *Plane*. For the pilot isn't blameworthy for diverting the plane, in the first place.[20]

But there is another problem with our second pass, which seems more fundamental. Consider this case:

> *Hold your breath*: With much effort you can hold your breath for one full minute. You know that, if you don't do that at a certain time, one person, X, will be in just a little pain for one second, and another person, Y, will suffer horribly for a month. You don't care about other people's wellbeing, so you don't hold your breath.

Intuitively, in this case you are blameworthy for choosing not to hold your breath. And this foreseeably led to X's being in a little pain for one second. However, intuitively, you're not blameworthy for this outcome: what you are blameworthy for is Y's being in a lot of pain (another foreseeable consequence of your choice), not X's being in a little pain.

Why aren't you blameworthy for X's being in a little pain, even if you are blameworthy for your behavior and your behavior led to that outcome in the expected way? Intuitively, this has to do with the *reason* you are blameworthy for your behavior in this case: given that the reason *isn't* that you knew that it would cause X a little pain, your blameworthiness for your behavior cannot then, as a result of

[20] On the other hand, a parallel principle about praiseworthiness doesn't entail that the pilot is praiseworthy for the deaths either, because the principle about praiseworthiness would be restricted to good outcomes only. One could also formulate a more neutral principle about a broader form of responsibility, one that didn't require the outcome to be good or bad. That principle presumably *would* entail that the pilot is responsible for the deaths in that sense, but this is arguably the right result.

DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767 *Crítica*

this mismatch, carry over to that particular outcome. In contrast, the reason you are blameworthy for your behavior is (at least partly) that it would cause Y a lot of pain; as a result, your blameworthiness for your behavior can—and in fact does—carry over to that outcome.

Again, it helps to cash this out in terms of grounding. On this view, an agent's responsibility for an outcome is grounded in, among other things, the agent's responsibility for some antecedent behavior. But now we can see that the responsibility for the antecedent behavior must, in turn, be grounded in a certain way—specifically, in a fact concerning the anticipated connection between the behavior and the outcome O (or an outcome of O's type).[21] This is, intuitively, what happens in *Hold your breath*: your behavior is blameworthy in that case *because*, or *by virtue of the fact that*, you knew that behaving in that way was likely to lead to some serious harm (the suffering by Y), and this is why you're blameworthy for that outcome.[22]

The moral seems to be this: for S to be blameworthy for O by virtue of B, S must be blameworthy for B, but S must be blameworthy for B for the relevant reasons, or by virtue of the relevant grounds. Those are grounds that "circle back" to O—not in a problematic or circular way, but just in a way that draws on the anticipated causing of O, or of an outcome of O's type. This suggests that outcome responsibility is, not just any kind of inherited responsibility, but responsibility that is inherited from a behavior only insofar as the anticipated connection to the outcome is already *part of the reason why* the agent is responsible for the behavior, in the first place.

We're now ready to formulate our third and final pass. Again, exactly how to formulate the additional requirement is debatable. (Should we require that the agent knew that the outcome would occur? Or that it was likely to occur? Should we merely require that the agent believed some such thing? Justifiably believed it?) Given that I'm only interested in the main structure of the view, in what follows I'll just use the expression "because of the relevant facts

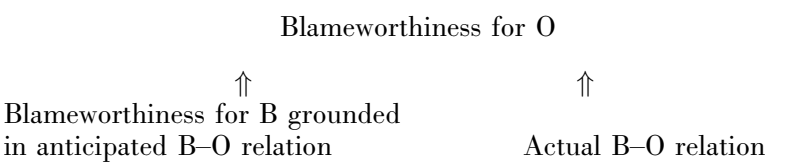
[21] Note that this is so even if the responsibility at issue is "basic" responsibility. Being "basically" responsible just means that it's not grounded in the agent's being *responsible for* some other antecedent behavior.

[22] Since we're assuming that outcome responsibility is a genuine phenomenon, we're setting aside views inspired by Kantian ideas about the "good will" being the only thing that matters and actual consequences not mattering (for a contemporary defense of this view, see Khoury 2018). But note that even views of this kind can accept that part of what makes us responsible for our actions are the *expected* consequences of those actions.

concerning the anticipated relation between B and O" to pick out the additional requirement, whatever it may be. This is what we get:

> (P3) S is blameworthy for O by virtue of B when S is blameworthy for B partly because of the relevant facts concerning the anticipated relation between B and O, and when B led to O, in roughly the expected way.

In terms of grounding, this means that non-basic responsibility is grounded in two main components:



The first component is what makes it a form of inherited responsibility. But, again, S's blameworthiness for O isn't just grounded in S's blameworthiness for B, for it's also grounded in the fact that S is blameworthy for B for the relevant reasons. As represented by the diagram, we can understand this in terms of grounding too, by appeal to "nested" grounding relations: S's blameworthiness for O is grounded in the fact that S's blameworthiness for B is grounded in the anticipated B–O relation. (Here I won't get into how exactly one should analyze the nested grounding relations, but I take it the idea is intuitive enough.) In turn, the second component of non-basic responsibility is of a quite different nature: it's the "metaphysical glue" that must in fact exist between B and O (be it causation or some form of non-causal consequence) for S to be responsible for O, and that also grounds S's responsibility for O.

This view offers a natural explanation of why you're blameworthy for the harm in our original case, *Button*. You're blameworthy for the harm, first, because you're blameworthy for pushing the button, and for the relevant reasons (because you could anticipate that it would cause harm); and, second, because pushing the button in fact resulted in the harm, in roughly the expected way.

Notice that both components are essential to this form of responsibility. On the one hand, if you hadn't been blameworthy for pushing the button (for the relevant reasons), you wouldn't have been blameworthy for the harm, even if pushing the button caused the harm.

Hence, the first component is needed. On the other hand, if you had been blameworthy (for the relevant reasons) for pushing the button but it hadn't caused the harm (in the expected way), you, again, wouldn't have been blameworthy for the harm. Hence, the second component is also needed.

I'll end by discussing applications of the account to two specific issues.

The first issue focuses on the role played by the second component (the "metaphysical glue"). The question is: given that non-basic responsibility is partly grounded in the relevant metaphysical glue, does this mean that one should always try to figure out who is responsible for an outcome by first figuring out if the relevant metaphysical glue obtains? In other words, do the grounding and epistemic orders march in lockstep, as far as this condition of non-basic responsibility is concerned?[23]

I'll argue that this isn't always the case. Typically, they do march in lockstep, but not always. Let's start with a case where they do. Imagine that pushing either of two buttons, button 1 or 2, would start an independent causal process leading to an explosion. Imagine that, wanting for the explosion to happen, you push button 1 and I push button 2. The explosion happens. Who is to blame for it?

We need more information to answer this question. Imagine that, upon further investigation, we discover that one process went to completion before the other, as in the following "preemption" scenario:

> *Preempting button*: Pushing button 1 resulted in the explosion before button 2 could do anything at all.

Now we know that your act in fact led to the explosion while mine did not. As a result, we can conclude that you're blameworthy for the explosion and I am not. Here we're using the empirical facts about the case to establish causal responsibility, and then we're using causal responsibility to establish moral responsibility for the outcome, by appeal to the conditions expressed in P3.

However, this strategy doesn't generalize to all cases. Consider, for example, this variant of the case:

> *Overdetermining buttons*: This time we discover that the two processes went to completion at exactly the same time.

[23] I discuss this in more detail in Sartorio (forthcoming b). See also Tierney 2023 for further discussion.

This is not a case of preemption but of "overdetermination". Who is responsible for the explosion in this case?

Clearly, we both are; to me this seems clearer than any claim about the causal structure of the case.[24] Metaphysicians would in fact disagree about its causal structure: "individualists" about causal overdetermination would claim that each of us individually is a cause, whereas "collectivists" would claim that a collective event (such as the sum of the two button-pushings) is a cause but that individually we are not causes (see Schaffer 2003 for discussion). Intuitively, our responsibility for the explosion in this case doesn't hinge at all on who is right about this (quite technical) metaphysical debate. Intuitively, "together" we made the explosion happen, in some way or other, and this is enough to make us responsible (either in an individualist or in a collectivist way). In other words, it's clearer *that* we are both responsible for the explosion than *how* exactly this is the case.

P3 has the flexibility to accommodate this. For all that's required for us to be responsible for the outcome, according to P3, is that there be some antecedent thing from which our responsibility could be inherited. So far, I've been assuming, for simplicity, that that thing is always an individual behavior. But it could also be a collective behavior. In fact, if collectivism is true, the relevant behavior would be a collective behavior (a single collective event like the sum of the two button-pushings): according to the collectivist story, both of us are partly responsible for that collective behavior, which resulted in the outcome, and that is why we are responsible for the outcome.[25]

In sum, although a certain kind of metaphysical glue grounds the non-basic responsibility of agents, as reflected by P3's second component, this doesn't mean that the best way to figure out who

[24] I take it this is intuitively clear, but here is an argument if you need one: surely, *somebody* is responsible for the explosion in this case, and given that the situation is perfectly symmetrical, if one is responsible then so is the other. What I admit might be a bit less clear in this case is the degree to which each of us is responsible for the outcome (although for an argument that we're just as responsible as we would have been if we had acted in isolation, see Zimmerman 1985).

[25] All that matters for my purposes here is that we are both responsible for the collective behavior, not how (but collectivism is compatible with our individual behaviors playing a role at that stage). Interestingly, Schaffer himself suggests that collectivism wouldn't allow us to blame agents for overdetermined outcomes (but only for lesser crimes like attempts, and this is partly why he is an individualist; see Schaffer 2003, n. 11). I think this is wrong: armed with P3, collectivists can blame agents for the full crime by blaming them for the antecedent collective behaviors.

DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767 *Crítica*

is responsible for an outcome is always to establish the underlying metaphysical facts first. Sometimes it's not, as when the moral facts themselves are clearer than the underlying metaphysical facts. In those cases, the grounding and epistemic orders do not march in lockstep.

The second issue that I'll discuss focuses mainly on the first component (the inherited nature of non-basic responsibility), and it concerns the problem of collective harms. The version of the problem I'm interested in goes as follows.[26] On the one hand, the contributions of individual agents can collectively "add up" to produce significantly harmful outcomes—for example, environmental disasters like pollution or climate change. Arguably, in at least some of those cases, we want to be able to blame at least some individuals (to at least some degree) for those outcomes. However, on the other hand, it's unclear how we could do this. For the individual contributions seem to be irrelevant to the occurrence of those outcomes, in two main ways: each contribution on its own is quite insignificant or minimal, and the outcomes are heavily overdetermined. For example, consider the contribution of a single individual to the pollution of the environment. That individual contribution seems negligible in that, on its own, it only contributes a tiny amount to the overall amount of pollutants and, without it, there would have been more than enough pollutants to ensure the same outcome. How can we be responsible for collective harms, then, if our individual contributions are negligible in these two ways?

There is a line of response to the problem, which I'll refer to as a "causal response", that in very general terms goes as follows.[27] In collective harm cases, despite our not making a difference, or a significant difference, to the outcomes in the ways specified above, we can—and in fact typically do—make a *causal contribution* to the occurrence of those outcomes by acting in the ways we do. As a result, we have a moral reason to refrain from acting in those ways and, under the right circumstances, we can be morally responsible (blameworthy, to at least some degree) for the outcome by so acting.

The motivation for causal responses seems clear enough: making a causal contribution to a harm is a way of "getting our hands dirty",

[26] This is related to the more classical version in terms of reasons to act (Parfit 1984, chapter 3) in that, if an individual is blameworthy for a collective harm, this is likely to be explained in terms of the individual's having acted in a certain way despite having had moral reasons not to act in that way.

[27] See Goldman 1999; Tuck 2008; Lee 2022; see also Kaiserman 2024 for related discussion.

at least to some extent, by being involved in the production of the harm, and this is something that we have at least some reason not to do. Thus, if we can show that our individual behaviors causally contribute to the occurrence of collective harms, despite not making a difference or a significant difference to those outcomes, this can go at least some way towards explaining our responsibility for those harms.

I'll argue that causal responses fail to provide a satisfying solution—or even the beginning of a solution—to the problem of collective harms, and that our discussion of non-basic responsibility can help us see why.[28] As we have seen, in order to establish an agent's blameworthiness for an outcome, it's not enough to establish the existence of a causal relation between the behavior and the outcome: we must also show that the agent is blameworthy for the behavior, and that this is so for reasons that have to do with the anticipated connection between the behavior and that outcome. So, the first thing to note is that, in a case of a collective harm, where we're wondering whether an individual is responsible for the harm, it's not enough to show that some behavior by an individual is a cause of the harm: we also need to show that the individual is blameworthy for that behavior, and for the relevant (outcome-related) reasons.

So far, this doesn't seem that bad. After all, nobody thinks that establishing causation *alone* is sufficient to establish moral responsibility, so it's not at all surprising that to establish responsibility for the outcome we need to do more than establish causation. But there is a second thing to note, which is more important. It is this: what's left to establish in these cases (again, that the individual is blameworthy for the behavior, and that this is so for the relevant, outcome-related reasons), is apparently *just as hard* as establishing that the individual is blameworthy for the outcome. It is just as hard because in both cases we face the same kinds of difficulties.

To illustrate, consider, again, an outcome like environmental pollution. Many individuals collectively contributed to this outcome over the years by each contributing some individually insignificant amounts of pollutants. Plus, the outcome was also heavily overdetermined (by multiple collections of individual contributions). Under these circumstances, it is hard to explain how each of those individuals could be blameworthy for the outcome. But (this is the rub) it

[28] Other philosophers have criticized causal approaches (see, e.g., Nefsky 2017; 2019). I take my objection to be importantly different from those other objections.

seems equally hard to explain how each of them could be blameworthy for their individual *behaviors*, or how they could be blameworthy for their individual behaviors *in light of* the anticipated connection between those behaviors and the outcome. And this is so for apparently the same kind of reason: because that connection is very feeble, or insignificant. As a result, causal responses to the problem of collective harms fail. They fail because establishing the existence of an individual causal connection doesn't make any significant progress towards explaining how individuals can be morally responsible for collective harms.

Our earlier discussion of the individualism/collectivism debate about overdetermination can be used to reinforce this argument. As we have seen, collective harms are overdetermined outcomes (heavily overdetermined outcomes, in fact). Proponents of causal responses are *individualists*: they claim that agents in collective harm cases contribute causally to those harms, qua individuals; moreover, it is *because* of those individual causal contributions that those agents can be morally responsible for the collective harms. However, as we have seen by appeal to a simpler overdetermination case, *Overdetermining buttons*, the truth of individualism is simply irrelevant to the moral responsibility of agents in overdetermination cases. For there is an alternative to individualism, collectivism, which could easily account for the agents' responsibility for the outcome, *if* the agents were indeed responsible. (In *Overdetermining buttons* the agents are, indeed, responsible for the outcome, and this is so *regardless* of whether individualism is true: the truth of individualism wouldn't make them any more responsible than the truth of collectivism would.) This suggests that individual causation versus collective causation is a red herring in the debate about responsibility for collective harms. It's a red herring in that debate because it's a red herring in the debate about overdetermined outcomes more generally.

I conclude that the solution to the problem of collective harms must be found, not in the metaphysics of causation, but elsewhere. On reflection, this shouldn't come as a big surprise. Wouldn't it be strange if the solution to the problem of collective harms laid in the metaphysics of causation—in particular, in the causal structure of overdetermination cases? Metaphysics is very cool and important indeed, but it seems delusional to think of it as having that kind of power.

Fortunately, there's also a more positive conclusion that I believe we can draw from all of this. It's that it pays to think more about

 DOI:https://doi.org/10.22201/iifs.18704905e.2026.1767

the nature of non-basic responsibility, and of how causalists can accommodate it. I've discussed two issues where this kind of reflection seems to have interesting consequences, but I invite the reader to think of others.[29]

REFERENCES

Beebee, Helen, 2004, "Causing and Nothingness", in John Collins, Ned Hall, and L.A. Paul (eds.), *Causation and Counterfactuals*, The MIT Press, Cambridge, Mass., pp. 291–308.

Bernstein, Sara, 2015, "The Metaphysics of Omissions", *Philosophy Compass*, vol. 10, no. 3, pp. 208–218.

Capes, Justin A., 2023, *Moral Responsibility and the Flicker of Freedom*, Oxford University Press, New York.

Chisholm, Roderick, 1964, "Human Freedom and the Self", in Robert Kane (ed.), *Free Will*, Blackwell, Malden, Mass.

Clarke, Randolph, 2015, "Abilities to Act", *Philosophy Compass*, vol. 10, no. 12, pp. 893–904.

Clarke, Randolph, 2009, "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism", *Mind*, vol. 118, no. 470, pp. 323–351.

Coates, D. Justin and Philip Swenson, 2013, "Reasons-Responsiveness and Degrees of Responsibility", *Philosophical Studies*, vol. 165, no. 2, pp. 629–645.

Cyr, Taylor W., 2017, "Semicompatibilism: No Ability to do Otherwise Required", *Philosophical Explorations*, vol. 20, no. 3, pp. 308–321.

Davidson, Donald, 1963, "Actions, Reasons, and Causes", *Journal of Philosophy*, vol. 60, pp. 685–700.

Dowe, Phil, 2001, "A Counterfactual Theory of Prevention and 'Causation' by Omission", *Australasian Journal of Philosophy*, vol. 79, no. 2, pp. 216–226.

Fara, Michael, 2008, "Masked Abilities and Compatibilism", *Mind*, vol. 117, no. 468, pp. 843–865.

Feinberg, Joel, 1970, "Sua Culpa", in *Doing and Deserving: Essays in the Theory of Responsibility*, Princeton University Press, Princeton, pp. 187–221.

[29] This essay is partly based on the Gaos lectures I gave at UNAM in March 2025. I am grateful to UNAM and the philosophers at the Instituto de Investigaciones Filosóficas for the invitation to give those lectures. It was a rich and stimulating experience that helped me rethink how the different pieces of my research project fit together, and this essay is a natural offshoot of the lectures and the reflection that followed them. Thanks to Santiago Echeverri and Miguel Angel Rotter for their editorial help with the preparation of the article and for their comments on an earlier draft.

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

Frankfurt, Harry G., 1978, "The Problem of Action", *American Philosophical Quarterly*, vol. 15, no. 2, pp. 157–162.

Frankfurt, Harry G., 1969, "Alternate Possibilities and Moral Responsibility", *The Journal of Philosophy*, vol. 66, no. 23, pp. 829–839.

Franklin, Christopher Evan, 2018, *A Minimal Libertarianism: Free Will and the Promise of Reduction*, Oxford University Press, New York.

Goldman, Alvin I., 1999, "Why Citizens Should Vote: A Causal Responsibility Approach", *Social Philosophy and Policy*, vol. 16, no. 2, pp. 201–217.

Griffith, Meghan, 2017, "Agent Causation", in Kevin Timpe, Meghan Griffith, and Neil Levy (ed.), *The Routledge Companion to Free Will*, Routledge, London and New York.

Jaster, Romy, 2022, "The Ability to Do Otherwise and the New Dispositionalism", *Inquiry*, vol. 65, no. 9, pp. 1149–1166.

Kaiserman, Alex, 2024, "Responsibility and Causation", in Maximilian Kiener (ed.), *The Routledge Handbook of Philosophy of Responsibility*, Routledge, New York, pp. 164–176.

Kaiserman, Alex, 2021, "Reasons-Sensitivity and Degrees of Free Will", *Philosophy and Phenomenological Research*, vol. 103, no. 3, pp. 687–709.

Kane, Robert, 1996, *The Significance of Free Will*, Oxford University Press, New York.

Kelley, Mikayla, 2024, "How to Perform a Nonbasic Action", *Noûs*, vol. 58, no. 1, pp. 106–125.

Kelley, Mikayla, (forthcoming), "A Control Theory of Action", *Australasian Journal of Philosophy*.

Khoury, Andrew C., 2018, "The Objects of Moral Responsibility", *Philosophical Studies*, vol. 175, no. 6, pp. 1357–1381.

Law, Andrew, 2022, "What does Indeterminism Offer to Agency?", *Australasian Journal of Philosophy*, vol. 100, no. 2, pp. 371–385.

Lee, Samuel, 2022, "Collective Actions, Individual Reasons, and the Metaphysics of Consequence", *Ethics*, vol. 133, no. 1, pp. 72–105.

Lewis, David, 1973, "Causation", *The Journal of Philosophy*, vol. 70, no. 17, pp. 556–567.

McKenna, Michael, 2013, "Reasons-Responsiveness, Agents and Mechanisms", in David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*, vol. 1, Oxford University Press, Oxford, pp. 151–183.

Mele, Alfred, 2017, *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will*, Oxford University Press, New York.

Metz, Joseph, (forthcoming), "Causalism and the Grounds of Agency", *Analysis*.

Nefsky, Julia, 2019, "Collective Harm and the Inefficacy Problem", *Philosophy Compass*, vol. 14, no. 4, e12587.

Nefsky, Julia, 2017, "How You Can Help, Without Making a Difference", *Philosophical Studies*, vol. 174, no. 11, pp. 2743–2767.

Nelkin, Dana Kay, 2025, "Moral Luck", *The Stanford Encyclopedia of Philosophy* (Fall 2025 Edition), E. Zalta and U. Nodelman (eds.), https://plato.stanford.edu/archives/fall2025/entries/moral-luck/

Nelkin, Dana Kay, 2016, "Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness", *Noûs*, vol. 50, no. 2, pp. 356–378.

O'Connor, Timothy, 2000, *Persons and Causes: The Metaphysics of Free Will*, Oxford University Press, New York.

Palmer, David, 2021, "Free Will and Control: A Noncausal Approach", *Synthese*, vol. 198, no. 10, pp. 10043–10062.

Parfit, Derek, 1984, *Reasons and Persons*, Clarendon Press, Oxford.

Sartorio, Carolina, 2023, *Causalism: Unifying Action and Free Action*, Oxford University Press, Oxford.

Sartorio, Carolina, 2022, "The Grounds of our Freedom", *Inquiry*, vol. 65, no. 10, pp. 1250–1268.

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

Sartorio, Carolina, (forthcoming a), "The Structure of Outcome Responsibility", in Mattias Gunnemyr, Rutger van Oeveren, and Jan Willem Wieland (ed.), *The Ethics of Inefficacy*, Routledge.

Sartorio, Carolina, (forthcoming b), "Responsibility, Causation, and the Cart-Horse Metaphor", in Christopher Lumer (ed.), *Retrospective Responsibility–Function and Conditions*, Springer.

Sartorio, Carolina, (forthcoming c), "Replies to Critics", *Analysis*.

Sartorio, Carolina, (Ms), "Basic and Non-Basic Responsibility: Against Unification".

Schaffer, Jonathan, 2003, "Overdetermining Causes", *Philosophical Studies*, vol. 114, no. 1–2, pp. 23–45.

Sehon, Scott, 2016, *Free Will and Action Explanation: A Non-Causal, Compatibilist Account*, Oxford University Press, Oxford.

Smith, Holly, 1983, "Culpable Ignorance", *The Philosophical Review*, vol. 92, no. 4, pp. 543–571.

Tierney, Hannah, 2023, "The Future of the Causal Quest", in Joseph Campbell, Kristin M. Mickelson, and V. Alan White (eds.), *A Companion to Free Will*, Blackwell, Hoboken, New Jersey.

Tierney, Hannah, 2019, "Quality of Reasons and Degrees of Moral Responsibility", *Australasian Journal of Philosophy*, vol. 97, no. 4, pp. 661–672.

Tuck, Richard, 2008, *Free Riding*, Harvard University Press, Cambridge, Mass.

Van Inwagen, Peter, 1983, *An Essay on Free Will*, Oxford University Press, Oxford.

Vetter, Barbara, 2019, "Are Abilities Dispositions?", *Synthese*, vol. 196, no. 196, pp. 201–220.
Vetter, Barbara and Romy Jaster, 2017, "Dispositional Accounts of Abilities", *Philosophy Compass*, vol. 12, no. 8, e12432.
Vihvelin, Kadri, 2013, *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*, Oxford University Press, Oxford.
Wallace, Robert H., 2023, "A Dilemma for Reductive Compatibilism", *Erkenntnis*, vol. 88, no. 7, pp. 2763–2785.
Whittle, Ann, 2010, "Dispositional Abilities", *Philosophers' Imprint*, vol. 10, pp. 1–23.
Wilson, George M., 1989, *The Intentionality of Human Action*, Stanford University Press, Stanford.
Zimmerman, Michael J., 1985, "Sharing Responsibility", *American Philosophical Quarterly*, vol. 22, no. 2, pp. 115–122.

Received: December 29, 2025; accepted: January 15, 2026.