Received: 23 December 2021 | Revised: 28 May 2022 | Accepted: 16 June 2022

DOI: 10.1111/phib.12275

ORIGINAL ARTICLE

ANALYTIC PHILOSOPHY WILEY

Against the inside out argument

Amy Seymour

Department of Philosophy, Fordham University, New York, USA

Correspondence
Amy Seymour, Department of Philosophy, Fordham University.
Email: aseymour3@fordham.edu

Funding information
Fordham University

Abstract

Bailey (2021) offers a clever argument for the compatibility of determinism and moral responsibility based on the nature of intrinsic intentions. The argument is mistaken on two counts. First, it is invalid. Second, even setting that first point aside, the argument proves too much: we would be blameworthy in paradigm cases of non-blameworthiness. I conclude that we cannot reason from intentions to responsibility solely from the "inside out"—our possessing a blameworthy intention cannot tell us whether this intention is also blameworthy in deterministic worlds.

1 | INTRODUCTION

We appear to be blameworthy for some of our intentions and mental states. This much seems uncontroversial, at least on the assumption that it's possible to be blameworthy for anything. But the conditions for being blameworthy are under dispute.[1] Bailey (2021) argues that if we are morally responsible for something mentally internal to us—say, a bad intention—then moral responsibility is compatible with determinism.[2]

Bailey's argument would be significant if successful: we would be able to reliably use our own internal states to prove compatibilism. Take a negative intention for which you are morally responsible. Bailey thinks having that blameworthy intention *itself* demonstrates compatibilism.

[1] See Clarke, R. K., McKenna, M., & Smith, A. M. (Eds., 2015) for an overview of this dispute. See particularly Zimmerman (2015) for distinctions regarding the kinds of responsibility at issue.

[2] In what follows, I will speak of "compatibilism about moral responsibility and determinism" as the abbreviated "compatibilism". Bailey thinks his argument might be extended to argue for compatibilism about freedom and determinism (p. 9) but leaves that particular discussion for the future. Since an extended version of the argument will suffer the same difficulties as the original, my argument will also apply to Inside Out arguments regarding freedom and determinism. I refute not just Bailey's argument, but the general strategy.

For intrinsic duplicates of you in deterministic worlds also have that blameworthy intention. So, you are responsible even if determinism is true.[3]

The "Inside Out" argument continues an important tradition of utilizing mental states to adjudicate freedom or responsibility. Some mental states and actions, such as deliberating about whether to perform an action, appear to provide some role in thinking we are free.[4] The Inside Out argument has even more power. The blameworthy intention is not merely evidence of some feature that demonstrates that responsibility is stable between indeterministic and deterministic worlds. Rather, the existence of the blameworthy intention itself gets the job done, as the blameworthiness for the intention is (assumed to be) intrinsic and thus stable across the considered contexts.

But the argument fails on two counts. First, it is invalid. Intrinsic duplicates can fail to manifest intrinsic properties in differing modal contexts—this difficulty underscores the debates about dispositions, especially in cases of masking or finking.[5] A glass may be intrinsically fragile and predisposed to shatter when struck but fail to have that property manifest via shattering in different scenarios, including differing laws of nature. And this is exactly what the incompatibilists can say about any supposed intrinsic properties of responsibility: they are dispositions, and only manifest in the appropriate (indeterministic) contexts. To fix the argument, one must provide a premise about intrinsic duplicates and moral responsibility that incompatibilists will deny.[6]

Second, the argument is too powerful: we, or our duplicates, are morally responsible in *any* scenario in which we have the particular regrettable intention. The Inside Out argument thus results in *responsibility explosion*—agents are ruled blameworthy in cases widely regarded to be non-controversial, paradigm cases where agents lack responsibility. If successful, the Inside Out argument would prove that individuals are responsible in cases which were our best shot at (agreed upon) non-blameworthiness.

An example of such a case involves Martian manipulation: there is widespread agreement that I am not blameworthy for mental states or actions that are a direct result of Martians manipulating me.[7] But take a morally responsible, intrinsic intention that I have. According to the

[3] Bailey does not think that simply any blameworthy internal intention will do the job (see his distinction on pp. 8–9 between "basic and internal" and "non-basic and external" cases). Bailey does think, however, that we have access to at least one negative intention with which we can run the above argument (see his p. 6). The difficulty in specifying which sorts of intention could be used for Bailey's reasoning is the central concern of this paper.

[4] See Kapitan (1986) and Markosian (1999) for the role of deliberation in indicating freedom for the compatibilist. See van Inwagen (2008) for an argument that our need for such deliberation would prevent even an omniscient being from knowing some future facts.

[5] See Johnston (1992) for the original use of masking, Fara (2005) for a wide variety of cases, and Lewis (1997) for a discussion of finking.

[6] The exact specification of such a premise will be tricky, since the Inside Out argument requires not only the *having* of the intrinsic property, but its manifestation in the appropriate instances. The incompatibilist will have a ready, motivated response to any additional premise attempting to block the dispositions response, so the specification of this premise need not be done here.

[7] The assumption that direct manipulation removes freedom and responsibility is so widespread that Mickelson (2017) simply labels this kind of case the "victim premise" of manipulation arguments (p. 170). Mickelson focuses on Pereboom (2001), who gives four cases regarding a Professor Plum, whose actions are controlled by neuroscientists. While Pereboom's arguments can also be put into an Inside-Out-style argument, the Martian manipulation case avoids reasonable concerns one might have with brain chips and intrinsic duplicates. Fischer (2004) notably disagrees with the general consensus, saying that the victim in the "victim premise" is indeed responsible. However, as will be noted later, Fischer's reasoning regarding Professor Plum's responsibility is currently unavailable to Bailey.

Inside Out argument, any intrinsic duplicate of mine will share the moral responsibility of this responsible intention. Since manipulation by the Martians is extrinsic, my intrinsic duplicates are culpable even when directly manipulated by the Martians. Something is wrong with this strategy of reasoning from the inside out.

I provide a diagnosis: a good duplicate is hard to find. Responsibility explosion demonstrates that not all seeming intrinsic properties are reliably transferrable across modal contexts. The Inside Out argument relies on intrinsic properties which are *transworld stable*—that is, properties which are impervious to external variation and retained by the individual, or their duplicates, across widely varying modal contexts.[8]

But this stability is not forthcoming. In Section 4, I show that no matter how we specify intrinsic properties in the search for transworld stability, we will undermine our ability to argue for compatibilism from the inside out. In Section 5, I demonstrate that only very particular kinds of mental states guarantee the transworld stability needed to reason from the inside out. Any being which thinks to themselves "I think; I exist" cannot be wrong about this judgment, come what may—reflection on that mental state self-supplies the reasons needed to accurately accept the target claim, even in wildly different modal contexts. But self-validating or self-supporting reasons are not available when internally considering our own responsibility: we can be wrong when considering whether we are blameworthy for our intentions.[9] My thinking that I am responsible does not guarantee that this is so in the actual world, let alone in counterfactual scenarios.

I conclude that the failure of "inside out" reasoning extends to any argument which runs solely from the existence of internal mental states to general conclusions about culpability, the laws, and possible worlds.[10] We are unfortunately not able to responsibly reason about responsibility from the inside out.

2 | FIXING THE INVALIDITY: MORAL RESPONSIBILITY AND DUPLICATES

Suppose that a person named Jo has a bad intention. Intentions, whatever else they are, appear internal to beings like Jo. Thus, Bailey assumes that such internal states or properties are intrinsic. He minimally defines moral responsibility as being an apt candidate for praise and blame. Determinism is the thesis that the state of the world at an instant together with the laws of nature

[8] Bailey is explicit in thinking that intrinsic properties generally provide this stability (see his pp. 3ff). As I demonstrate in Section 2, it is not obvious that all intrinsic properties have transworld stability.

[9] I can, perhaps, be reliable regarding other negative evaluations of my mental states or intentions. But as Smart (1961) points out, grading and evaluating the overall goodness of my intentions does not imply that I am responsible or appropriately blameworthy. My dispraise of my mental state may be more akin to my judgment of the quality of, say, good and bad apples (see Smart's pp. 303ff). This point is explored in Section 3.

[10] The Consequence Argument, for example, does not reason solely from internal mental states; it focuses on ability and looks directly at how ability relates to the laws of nature and states of the world at an instant (see especially chapter 3 of van Inwagen (1983)). Note, too, that Strawson's (1962) rejection of the classical debate in favor of the reactive attitudes also does not rely solely on internal states. Our actual practices of praise, blame, and the like—practices which do not solely rely on a single individual's mental state—matter.

entails only one physically possible future.[11] Since determinism is a global thesis, it is not an apt candidate for being internal or intrinsic to beings like Jo. Bailey's Inside Out argument proceeds as follows, in Bailey's words (p. 2):

1. *Being morally responsible for the bad intention* is intrinsic to Jo.
2. *Whether determinism is true* is extrinsic to Jo (i.e., either *being such that determinism is true* is extrinsic to Jo or *being such that determinism is false* is extrinsic to Jo).
3. If *whether determinism is true* is extrinsic to Jo, then Jo has an intrinsic duplicate at a deterministic world.
4. Therefore, Jo has an intrinsic duplicate at a deterministic world (from 2 to 3).
5. Therefore, Jo has a morally responsible duplicate at a deterministic world (from 1 to 4).
6. Therefore, possibly: someone is morally responsible and determinism is true (from 5).

It is not as clear as one might hope that moral responsibility—even if only for intentions—is an intrinsic property, especially one which allows for the mix-and-match modal approach for which Bailey advocates. Suppose, for now, that moral responsibility can be intrinsic. Bailey thinks the above argument can be easily applied, as intrinsic properties are "impervious to external variation", such that one can "...shift as you may, you will not change the bad intention itself. Nor will you change the appropriate stance toward Jo and her bad intention" (3). "Merely dropping Jo into different surroundings would not do the trick" (2). Intrinsic properties, after all, are about the relevant individual and not the world at large.

But many of my intrinsic properties—for example, having my particular shape—depend on the cooperation of many things extrinsic to me and cannot withstand modally shifting as I may.[12] I have a shape and a property like my shape is a paradigm case of an intrinsic property. However, I would not have this shape in a world with radically different physical laws.[13] My shape depends on certain gravitational laws and so forth and, without those features of the world, any intrinsic duplicate of mine would not retain the specific structure I have.

One response is that any being who does not share my identical shape cannot be a candidate for my intrinsic duplicate. But this response undercuts the motivation for premise 3 and we will return to it shortly.

One way to test for intrinsic properties is to use the "lonely world" test (Lewis, 1983). What properties would be retained by an intrinsic duplicate of mine in a lonely universe consisting of only my duplicate? Such a world cannot be *entirely* lonely; we'll need some laws and anything else my existence depends on, lest my "duplicate" be a lifeless, amorphous blob—or, at least, not long lived. If we understand intrinsic properties dispositionally, a case of intrinsic duplicates with radically different shapes (due to radically different laws) appears possible.

[11] There are other (logically equivalent) ways of specifying this thesis, including definition without explicit use of the laws. Take any two distinct times—these times will mutually entail every other time (see van Inwagen (1983)).

[12] If thinking is necessarily a diachronic relation or activity, then the intrinsic property of *my being a thinking thing* at this moment requires external support (at least, external to this particular time).

[13] My intrinsic duplicate need not survive long, if at all, in such a world. We are able to test for intrinsic duplicates using either an instantaneous time slice or by assuming endurantism. One could try to resist this argument by insisting that any true intrinsic duplicate of me must be a perdurantist spacetime worm, constituted of all and only spacetime slices of every moment at which I actually exist. This insistence, however, makes identifying an intrinsic duplicate difficult and gives us reason to think that contexts in which I have an intrinsic duplicate are much more limited—too limited for the Inside Out argument to be of use. See Section 4 for more on this point.

Lewis himself thinks of intrinsic properties in dispositional terms; we cannot reason solely from intrinsic duplicates to the laws or vice versa. According to Lewis, "if two things are exact intrinsic duplicates (and *if they are subject to the same laws of nature*) then they are disposed to be alike" (1997, p. 147, emphasis mine). Intrinsic duplicates will not necessarily be identical if the circumstances vary enough. This is a lesson from debates about dispositions. Take a vase with the intrinsic property of *fragility* (see Johnston (1992) for such cases). The vase is thus prone to shatter when struck. But consider an intrinsic duplicate of this vase in a world with a protective wizard, who casts a spell to prevent the vase from shattering when struck. Or consider a duplicate of the vase in a world with highly different laws of nature. These duplicates will not shatter, though our vase will. Importantly, differing physical laws can result in intrinsic duplicates which are not alike—even if they have all the same properties, these properties do not manifest in every situation. On the assumption that laws of nature are contingent, fragile vases have intrinsic duplicates which will not shatter.

We are now able to see a fatal flaw in the Inside Out argument—it is invalid. The move from premise 4 to premise 5 does not follow. Jo can have an intrinsic duplicate at a deterministic world (premise 4) but fail to have a morally responsible duplicate at that world (premise 5). If moral responsibility is intrinsic, we have no reason to think it is highly different from paradigm properties like my shape or tricky cases like fragility. Perhaps responsibility is dispositional and only appears in the right contexts, with the right laws.

Alternatively, perhaps intrinsic properties are not dispositional but rather are fine-grained. For example, the relevant property of our vase is not *fragility*, but perhaps *fragile in laws of nature L* or *fragile when not protected by a wizard*. This, too, makes it so we cannot validly move from premise 4 to premise 5, since it is not immediately unreasonable to think (as per the incompatibilist) that the appropriately fine-grained property is not *being morally responsible for the bad intention* but something like *being morally responsible for the bad intention due to the right conditions*. And it does not take much to provide motivated incompatibilist intuitions for "the right conditions", giving us something like *being morally responsible for the bad intention only in indeterministic contexts*. Note that this intrinsic property, if it indeed is one, is transworld—all of my intrinsic duplicates have it, even if they are in deterministic worlds. But the incompatibilist can simply say that my duplicates in deterministic worlds are not morally responsible since they are never in indeterministic contexts.

Jo's intrinsic duplicates can exist in deterministic worlds and fail to be morally responsible. Variation amongst intrinsic duplicates shows that the Inside Out argument fails, for we no longer have reason to think that intrinsic duplicates share their moral properties (or the relevant manifestations) across worlds. Without knowing more about intrinsic properties, we cannot move from premise 4 to 5. Given the debates about the nature of intrinsicality, it does not appear we have a ready conception available which would both secure the validity of the Inside Out argument and convince a neutral observer. And note that this point holds regardless of our particular conception about the nature of the laws.[14]

[14] One could argue against the point on Humean grounds. Suppose the laws are simply generalizations regarding constant conjunctions of events and so forth. Then it seems possible for me to have an intrinsic duplicate—with my shape and moral properties—in worlds with fundamental laws radically different than ours, as long as these laws have built into them an exception *just for my duplicate at the relevant instant*, which would allow for the same shape, moral responsibility, et cetera. This response, however, is fundamentally ad hoc and seems much more questionable than the discussion of laws and universals Bailey objects to on p. 4. And this sort of response might give us reason to think that moral responsibility would not be retained by my duplicate in this odd sort of world due to lack of relevant ability. Bailey anticipates the latter difficulty in his footnote 8. To the extent that one divorces ability from responsibility, one will find themselves simply biting the bullet of the argument about Martian manipulation that is to come.

The incompatibilist thinks that responsibility, intrinsic or not, partially depends on the laws. It is reasonable to think that responsibility requires certain conditions. So, it is no surprise that the incompatibilist thinks any appropriate moral counterpart of me—let alone an intrinsic duplicate—will require similar laws. To assume otherwise with this modal mix-and-match strategy begs the question against the incompatibilist.

Let me put the idea a bit differently. If my having my specific shape requires the cooperation of fundamental physical laws, it seems no stretch to suppose my being morally responsible requires such cooperation from the laws as well, and likely certain historical tracing conditions. Holding this position does not mean one thinks things are (in Bailey's words) "maximally modally fragile" (p. 4), but rather it respects the kind of thing that I am and what my intrinsic properties ontologically require.

The only way to make the Inside Out argument valid without adding a premise is if the laws are not contingent. Then, intrinsic duplicates will not vary across worlds. But contingency of the laws is necessary for the success of the Inside Out argument. If the laws are necessary, the compatibilist and incompatibilist are at an impasse which cannot be solved by the Inside Out argument. Someone is right, but the argument does not provide the resources to determine who.[15] (Necessity in general undermines motivations for premise three. A theist who believes they have the intrinsic property *being made in the image of God* will not think they have a qualitative duplicate who has this property in a world in which God does not exist.)

If the Inside Out argument is to succeed, we need transworld-stable intrinsic properties, which are impervious to external variation. There are motivated ways to repair the argument along these lines. According to Mark Johnston, moral status should follow intrinsic duplicates across worlds—moral status should be modally stable.[16] Applying Johnston's general reasoning about moral status to responsibility, we get the following addition to the Inside Out argument:

> *Responsibility Duplication:* For all possible worlds *w* and *v*, times *t* and *t**, and possible objects *x* and *y*, if *x* in *w* at *t* is an intrinsic duplicate of *y* in *v* at *t**, then *x* has moral responsibility in *w* at *t* iff *y* has moral responsibility in *v* at *t**.

This is a good first step, but more will need to be done to block concerns about dispositions. Intrinsic properties, at least more complicated ones such as *being morally responsible for a bad intention*, might not be easily retained or manifested cross-world.[17] How intrinsic properties behave in differing worlds is already controversial; we cannot easily utilize them to settle further controversial

[15] Many incompatibilists will think the laws are necessary. Thus, any reasoning from the Inside Out via necessary laws would give us reason to reject premise three.

[16] Johnston (2016a) and Johnston (2016b) is concerned with the moral status of intrinsic duplicates of spacetime worms. If *being a person* is a maximal intrinsic property and perdurantism is true (and so persons are maximal fusions of instantaneous temporal parts), intrinsic duplicates of persons will not always be persons. And if only persons, not parts, have moral status, then some intrinsic duplicates of persons lack moral status because they are not persons. Rather, these "personites" are parts of a larger spacetime worm. See Johnston (2016a) for an introduction to this personite problem, and Johnston (2016b) and Kaiserman (2019) for debate regarding whether this problem undermines four-dimensionalist views of persistence through time.

[17] We can raise similar concerns for intrinsic properties like *being a temporal part of a temporally continuing object*. While this property depends on things external to the instantaneous time slice, the property appears to be intrinsic; it is about the time slice itself. It tells us that the time slice is a part of something. And the property is modally fragile because we can easily find worlds in which an otherwise duplicate of the time slice lacks the property.

issues about responsibility. It will be difficult to convince a neutral observer since what is happening extrinsically can matter for intrinsic properties. But since the argument can be made valid, it is worth looking at the argument and intrinsic properties in general. Even granting the validity of the argument, trouble awaits us.

3 | RESPONSIBILITY EXPLOSION

Let us assume, for now, that Bailey is correct about the imperviousness of intrinsic properties and their ease of modal use. Here is one way to "shift as I may": There is an intrinsic duplicate of Jo who exists in a world where her bad intention is the direct result of Martian manipulation. Again, this sort of manipulation case is taken to be a paradigm example of a case in which agents lack moral responsibility.[18] Bailey has provided us reason to think that the entirety of the intrinsic property—that is, *being morally responsible for the bad intention*—is shared by Jo's duplicate. After all, the bad intention is internal to Jo. And we can even specify that this intention has the same internal feel and flavor, and plays the same roles, in the duplicate's mental life: it is an unkind intention to which she might readily assent once it is present. It certainly *feels* to the duplicate, from the inside, that she is responsible and the intention is located in the same place in her mind. (We can specify that the manipulation took place at a time outside of the temporal boundaries of consideration for our duplicate, and that this is a true mental duplicate: these Martians are capable of particularly sophisticated manipulation.[19]).

Bailey has said we can drop a duplicate of Jo into different surroundings and she (or her duplicate) will retain responsibility. In Bailey's own words: "[I]f you wanted to change the fact that Jo is blameworthy for her unkind intention, you would need to change something about Jo herself, by eliminating that intention altogether, somehow shifting it around in her mind, or even removing Jo from the picture entirely. Merely dropping Jo into different surroundings would not do the trick" (2). This reasoning about what is intrinsic has given us reason to think we have a duplicate of Jo in such a world. We've specified that we have not changed anything about Jo herself; all we have changed is the causal story for how Jo got that way. At minimum, the reasoning Bailey provides for thinking that *being morally responsible for a bad intention* is stable across worlds in the Inside Out argument is equally applicable here.

Whether the bad intention exists as a direct result of Martian manipulation, however, is extrinsic to Jo and her duplicates. For *being such that an intention results as a direct result of Martian manipulation* is extrinsic to Jo: it is part of the world more generally, and not a part of or about Jo herself. It is not, in Bailey's words, within her "internal boundaries" (p. 2). At minimum, the reasoning Bailey provides for thinking that *being such that determinism is true* is extrinsic is equally applicable here.

[18] Again, Fischer (2004) is a notable voice of disagreement. Fischer thinks the professor in Pereboom's (2001) cases is responsible even when directly manipulated. However, Fischer motivates this position by drawing a sharp distinction between moral responsibility and blameworthiness (or praiseworthiness). Since Bailey identifies moral responsibility with being an apt candidate of praise or blame, this way of escape is unavailable to him.

[19] You may be concerned that we cannot restrict times in such a way—a *true* duplicate will share other features with Jo that will prevent the following Martian manipulation case. I consider and answer this objection in Section 4. Preventing this sort of temporal restriction easily provides motivated reasons to reject either Bailey's premise 1 or premise 3.

We can follow Bailey's general reasoning to formulate the following argument:

MM1. *Being morally responsible for the bad intention* is intrinsic to Jo.
MM2. *Whether Jo's intention exists as a result of direct Martian manipulation* is extrinsic to Jo (i.e., either *being such that the intention exists as a direct result of Martian manipulation is true* is extrinsic to Jo or *being such that the intention exists as a direct result of Martian manipulation is false* is extrinsic to Jo).
MM3. If *whether Jo's intention exists as a result of direct Martian manipulation is true* is extrinsic to Jo, then Jo has an intrinsic duplicate at a world in which the intention exists as a direct result of Martian manipulation.
MM4. Therefore, Jo has an intrinsic duplicate at a world in which the intention exists as a direct result of Martian manipulation (from MM2 to MM3).
MM5. Therefore, Jo has a morally responsible duplicate at a world in which the intention exists as a direct result of Martian manipulation (from MM1 to MM4).
MM6. Therefore, possibly: someone is morally responsible for an intention formed as a direct result of Martian manipulation (from MM5).

Thus, the success of the Inside Out argument results in responsibility explosion: agents are responsible even in cases in which we had previously widely agreed that they were not. In fact, Bailey appears *committed* to the moral responsibility of a duplicate who has her bad intention as the result of direct Martian manipulation: “Would every intrinsic duplicate of that paradigm also be morally responsible, even if in a different environment? I think so. And I think reflection here supports premise one, even if various recondite theories of moral responsibility tell against it” (6).

But thinking that responsibility is not present in Martian manipulation cases seems anything but recondite.

The morally explosive result—responsibility in Martian manipulation worlds—is, in part, due to the ease-of-use for which Bailey advocates. He writes, “Jo's intrinsic duplicates will vary across many dimensions, but all of them will harbor some particularly unkind intention and will therefore be *equally apt* candidates for blame” (p. 2, emphasis mine). The Inside Out argument is supposed to be metaphysically easy; the possession of the unkind intention *itself* is what provides not only the responsibility, but the same level of responsibility, for intrinsic duplicates across a wide variety of contexts. By Bailey's reasoning, having the unkind intention is not only enough for Jo's duplicate to be blameworthy, but *equally* morally blameworthy. We've dropped Jo's duplicate into different surroundings and found morally explosive consequences.

We can run this sort of parody argument for many cases of supposed paradigm non-responsibility, including ones which do not rely on causal laws at all. So, we can avoid altogether concerns the proponent of the Inside Out argument might have regarding causal history and the laws. For instance, we can run an Inside-Out-style argument to conclude that we are responsible even in fatalist worlds: after all, *whether fatalism is true* is extrinsic to Jo and her duplicates.[20]

[20] The argument is as follows:

F1. *Being morally responsible for the bad intention* is intrinsic to Jo.
F2. *Whether fatalism is true* is extrinsic to Jo (i.e., either *being such that fatalism is true* is extrinsic to Jo or *being such that fatalism is false* is extrinsic to Jo).
F3. If *whether fatalism is true* is extrinsic to Jo, then Jo has an intrinsic duplicate at a fatalistic world.
F4. Therefore, Jo has an intrinsic duplicate at a fatalistic world (from F2 to F3).
F5. Therefore, Jo has a morally responsible duplicate at a fatalistic world (from F1 to F4).
F6. Therefore, possibly: someone is morally responsible and fatalism is true (from F5).

According to fatalism, it is a logical consequence that agents cannot do otherwise than they in fact do. In fatalist worlds, then, beings like Jo appear to be as responsible for their actions as other natural phenomena, such as hurricanes.[21] And since hurricanes are not appropriate targets of praise or blame, neither, it seems, are beings in fatalist worlds. If responsibility requires ability, those in fatalist worlds are not responsible by definition.

Even the most ardent deflationist in the responsibility debates will admit that there are some cases in which a person is not responsible (see Strawson, 1962). But if all that matters is simply *having* the bad intention, Inside Out arguments return the verdict that they are morally responsible. Again, responsibility explodes.

It is non-controversial that there are some intrinsic properties for which we are not responsible. For example, as Smart (1961) points out, I am not responsible for the shape of my nose. It is also non-controversial that we are able to morally grade the intentions of beings like Jo or her duplicates in Martian manipulation or fatalist worlds, similar to how we might grade the goodness of apples. But our ability to morally grade or dispraise an individual does not imply that the individual is blameworthy or responsible (Smart, pp. 303ff). Even mere ascriptions of responsibility need much more, perhaps a pragmatic justification (p. 302). Beings in Martian manipulation worlds and fatalist worlds are as responsible for their intentions as they are for the shape of their nose—that is, in no way.

To stop the responsibility explosion, we must do one of three things: (a) deny that responsibility is an intrinsic property (at least in the above parody arguments), (b) deny that intrinsic duplicates retain all intrinsic properties in every world in which they exist, or (c) deny that the individual in Martian manipulation and fatalist worlds *is* truly an intrinsic duplicate.

Here's the trouble: each of (a–c) refutes or undermines one of the premises of the Inside Out argument. Denying that responsibility is intrinsic (option (a)) is to reject the Inside-Out-style arguments at their outset by denying premise 1.[22] If we deny that intrinsic duplicates always retain all intrinsic properties in every world in which they exist (option (b)), then we again rest on the reasoning regarding duplicates that demonstrated the argument to be invalid. That leaves (c), the "no true duplicate" response: the individual in the Martian manipulation and fatalist worlds is not an intrinsic duplicate. Depending on the reason for denying that this is a true duplicate, we either deny premise 2 or 3: either something seemingly extrinsic is not truly extrinsic or we cannot have intrinsic duplicates in just any extrinsically varying scenarios.

[21] Being currently morally unreachable need not entail that duplicates sharing this status need have arrived there in the same way. Differing roads may lead to an internal state that is situated in the same place and plays the same role in one's mental economy. One person may have willfully insulated themselves from counterevidence. Another may have arrived there by direct Martian manipulation. It seems not at all implausible that the former but not the latter person is morally blameworthy.

[22] There might be a reason to reject the intrinsicality of the responsibility in the Martian manipulation case but not in Bailey's original Inside Out argument. To make this move, one must supply motivated reason for thinking there is a difference between cases. I do not think such a reason is forthcoming; Bailey underspecifies the case and I suspect any appropriate specification will either be one which the incompatibilist can also use or which results in responsibility explosion. So again, the argument either fails or proves too much. I consider a possible, motivated way of distinguishing these cases in Section 5. I argue it is not available to the proponent of the Inside Out argument and show that the response at best renders Inside Out arguments inert.

4 | A GOOD DUPLICATE IS HARD TO FIND

The proponent of the Inside Out argument might offer an objection: since the Martian manipulation *directly* brought about the intention, the manipulation cannot be regarded as appropriately extrinsic. The Martian manipulation argument does not have an intrinsic duplicate of Jo. Possession of the bad intention itself is not the relevant feature in our cross-world search for duplicates, but rather the possession of the *blameworthy* intention. To provide cases in which Jo gained the property in a different way (or there is some other relevant external difference) somehow shifts the intention around in her mind in a way that negates responsibility. The causal history can matter, either for whether there is an intrinsic duplicate or the very nature of the intrinsic properties themselves.

But note that "the causal story matters" is exactly what the incompatibilist will say about cases in which determinism is true. The incompatibilist can say that Jo does not have an intrinsic duplicate in deterministic worlds because the causal history matters—either the causal history is related to the intrinsic property itself or to whether Jo can have an intrinsic duplicate in such worlds. A "no true duplicate" response to the Martian manipulation argument gives us a good reason to think we aren't able to tell whether the individual in a deterministic world truly is an intrinsic duplicate, and thus whether there is moral responsibility in that world.

The "no true duplicate" response renders the Inside Out argument useless. It also seems contrary to the way Bailey wants to apply the argument. He says, "shift as you may, you will not change the bad intention itself"—and that the relevant intrinsic properties are "impervious to external variation" (p. 3). But the "shift as you may" permissiveness allows for Martian manipulation cases and the like. Imperviousness to variation *should* allow for easy cross-world testing for duplicates. But to the extent that it does, we land immediately in cases which most take to be paradigm cases of non-responsibility.

To the extent that we are not allowed to easily shift things around, the Inside Out argument loses its usefulness. To block responsibility explosion, one must say that would-be intrinsic duplicates do not share the property of moral responsibility in manipulation cases. That is, there are cases in which it might *appear* that there is an intrinsic duplicate, but in fact there is not—the intrinsic property has significantly altered or disappeared in the Martian case. To embrace the "no true duplicate" response is thus to abandon the argument. Even if we hold fast to the added (necessary) premise about moral responsibility and intrinsic duplicates, we are unable to tell in individual cases (at least, in the controversial ones) whether the object we are considering is indeed an intrinsic duplicate and thus responsible.

But we need to know *whom to look for* in other worlds when we are searching for intrinsic duplicates and applying various modal tests with them, especially if our internal states are supposed to give us any guidance about whether we are responsible should determinism be true. When have we located a true duplicate? What scenarios license "inside out" reasoning?

Bailey thinks reflection tells in favor of cases of intrinsic responsibility and intends for the Inside Out argument to be more readily used than what he considers more "recondite" theories of moral responsibility. But reflection on the nature of intrinsic properties and what they entail can be recondite indeed. We must specify, as much as possible, what sort of intrinsic property is relevant as well as criteria for candidate duplicates.

Bailey presents his argument in general terms, not specifying whether the relevant intrinsic moral properties are time-indexed or apply to Jo's (or her duplicate's) total world history (as he notes on p. 6). But this generality is a cheat: it makes it seem as if the proponent of the Inside Out argument can both block important tracing condition concerns about how we got those intrinsic properties (such as intentions) in the first place *and* consider Jo's intentions at a single time. Is the relevant intrinsic duplicate a time-slice or someone who shares (enough of) my world history?

Either option spells trouble for the Inside Out argument. Suppose the duplicate is a time-slice.[23] Then the causal history of Jo's having the intention is not relevant when searching for intrinsic duplicates (at least, in so far as we want to uphold premise two of Inside Out arguments). This means that there are intrinsic duplicates of Jo who have the bad intention *solely because* of direct, immediate manipulation from Martians. Responsibility explodes.

Suppose, then, that intrinsic duplicates merely share enough of the relevant world history. Here we encounter a problem regarding how to specify this relevance. (And we inherit a skeptical problem, too, as we might wonder when we have enough of a shared history or other features: even a committed compatibilist finds herself in a position where the nature of duplicates is unclear.) We cannot simply specify that the relevant world history is the total history of the individual (as Bailey suggests on p. 6). To do so would be to give the incompatibilist everything she needs, as no intrinsic duplicate of me can exist in a deterministic world (assuming this world is indeterministic) and vice versa. A being sharing my total world history will also share the causal laws.[24] So, premise three is false.

The proponent of the Inside Out argument must specify how much of the total world history to consider. They will need enough shared history to resist the allowance of things like manipulation cases, but not so much that duplicates across deterministic and indeterministic worlds are ruled out. This specification does not appear forthcoming, since we may insert Martian manipulation at any point in the causal history. Bailey put forward an argument by which we are supposed to reason, from the inside, about whether we are responsible; "no true duplicate" responses take away our ability to do so and thus remove the power of the argument.[25]

[23] Time-slices are used for ease of explication. The endurantist can simply specify the intrinsic properties to be considered at a particular time.

[24] Bailey tries to resist this in discussions of what the laws are, and their being extrinsic to persons in the worlds. (See his p. 4.) But this result follows even if we are not considering the laws per se. While determinism is often specified in terms of the state of the world at an instant and the laws of nature, the thesis can be just as aptly described using entailment from times alone (see Bailey's "*Determinism* is true just if that there is, at any time, exactly one physically possible future" (p. 1, emphasis his), along with van Inwagen (1983)). Take any two distinct states of the world at an instant (say, t_1 and t_5). Suppose determinism is true. Then t_1 and t_5 will together entail exactly what happens at every other time. Now consider a person, Flo, who exists in a deterministic world. Flo's total history is deterministic: any two distinct times which are part of her history together entail every other time at which she exists. Now consider Jo, who we'll specify exists in an indeterministic world. Jo's total world history is not deterministic. That is, any two distinct arbitrarily chosen times in her history will *not* together entail every other time in her history. Jo and Flo's total world histories thus have different properties or relations. For instance, Flo's world history has entailment relations that Jo's does not. We can specify these differing relations in terms of (seeming) intrinsic properties. Flo's total history has the property *being deterministic* while Jo's has the property *being indeterministic*. And note that these properties are great candidates for being intrinsic properties, unlike the general properties of *being such that determinism is true* or *being such that determinism is false* – for they are properties about the *individuals'* total histories, rather than the world or the laws writ large. So, perhaps we have reason to deny Bailey's premise 2 as well. But at minimum it follows that there are no intrinsic duplicates between deterministic and indeterministic worlds if duplicates share total world histories.

[25] Note that this response does not imply that we can never tell from the inside—say, from introspective reasoning—whether we actually are responsible for an intention we have. Rather, it's that we cannot take our introspective reasoning about responsibility for our intentions and apply it across worlds in cases where it is unclear at best whether the necessary conditions for responsibility have been met.

5 | REASONING FROM THE INSIDE OUT

One response, in light of the above, is to conclude that responsibility is extrinsic. Bailey thinks that easy, non-recondite reflection supports the assumption that *being morally responsible for the bad intention* is intrinsic to Jo. We have enough epistemic access to such properties, and their intrinsic nature, to license reasoning across possible worlds and develop a theory from consideration of such cases.

Let us consider, then, what appears to be an analogous case and see what lessons we can draw from our reflections. Consider a belief I have formed on the basis of sense perception, such as *there is a tree outside*. It seems reasonable to think holding this belief results in my having an intrinsic property: *believing that there is a tree outside*. Further, let us assume in this case that my belief counts as knowledge. Is *knowing that there is a tree outside* an intrinsic property of mine? Following Bailey's reasoning, it's not obvious that it is not. I am the knower; the knowledge is, in some sense, *about me* (and my mental state). So, let us suppose for now that my *knowing that there is a tree outside* is intrinsic to me.

Whether I exist in a world where my perceptual seemings or sense impressions are veridical, on the other hand, is not internal nor intrinsic to me. Consider a world in which the only beings that exist are myself (or my duplicate) and an evil demon hell-bent on deceiving me about whether an external world exists. In this demon world, there are no trees despite how much it might seem to me (or my duplicate) that there are. But *being such that a belief results from an evil demon* is extrinsic to me and my duplicates: it is part of the world more generally, and not a part of or about myself. It is not within my internal boundaries.

We can now run another parody argument:

ED1. *Knowing there is a tree outside* is intrinsic to me.
ED2. *Whether I exist in an evil demon world is true* is extrinsic to me (i.e., either *being such that I am in an evil demon world is true* is extrinsic to me or *being such that I am in an evil demon world is false* is extrinsic to me).
ED3. If *whether I exist in an evil demon world is true* is extrinsic to me, then I have an intrinsic duplicate at an evil demon world.
ED4. Therefore, I have an intrinsic duplicate at an evil demon world (from ED2 to ED3).
ED5. Therefore, I have a duplicate who knows there is a tree outside at an evil demon world (from ED1 to ED4).
ED6. Therefore, possibly: someone knows there are trees outside and there are no trees (from ED6).

What's going on here? A few lessons are immediately apparent. First, we have picked the wrong candidate property for "inside out" reasoning. Either *knowing there is a tree outside* is not an intrinsic property or it is not one which we are able to reliably use in this sort of cross-world reasoning. We cannot reliably use the property in cross-world reasoning because it is not transworld stable: general conditions of the world either factor into the property itself or make it so we cannot reliably use the property to locate duplicates of mine in vastly different modal contexts. Either way, I cannot reliably shift as I may and find the appropriate duplicate. So, either ED1, ED2, or ED3 is false.

The evil demon argument should be resisted and it's instructive to consider how to do so. First, one could give the "no true intrinsic duplicate" response: intrinsic duplicates of me will not be present in evil demon worlds since my belief is formed on the basis of my senses and is thus essentially a perceptual belief. So, the intrinsic property in connection with my intrinsic belief that *there is a tree outside* is really the property *having a perceptual belief that there is a tree outside* (and thus the intrinsic knowledge property is something like *perceptually knowing that there is a tree outside*). Since there is no genuine or veridical sense perception in evil demon worlds, I cannot have an intrinsic duplicate in such worlds. The property in question is not fully saturated; properly doing so will allow us to escape all parody arguments.

This move renders the Evil Demon Inside Out argument useless. If the relevant sort of intrinsic property is, fully saturated, *having a perceptual belief that there is a tree outside*, then premise ED3 of the evil demon argument is false.[26] True intrinsic duplicates are incredibly hard to find, as individuals who lack knowledge (even those in nearby worlds in which, say, safety or sensitivity fails) are not my duplicates regardless of how much they otherwise are like me. All individuals in the closest possible worlds in which knowledge fails are not intrinsic duplicates.

Now consider Bailey's original Inside Out argument. If the intrinsic properties necessary for the Inside Out argument must be detailed or saturated, then we have no reason to think that the property under consideration in the first premise of the Inside Out argument—*being morally responsible for the bad intention*—is fully saturated.

To avoid Martian manipulation and fatalist cases, the proponent of a "no true duplicate" response must say that the conditions for forming the intention are somehow included in the intrinsic property itself. It is unclear how these properties or conditions could be formulated to block the parody arguments without giving the incompatibilist everything she needs to resist the original argument. Perhaps extrinsic facts matter in the formation of the intrinsic mental state, or for the location or role the mental state plays in one's mental economy. Depending on the particular role that such seeming extrinsic facts play, we will reject either premise 1, 2, or 3. The incompatibilist can insist that the property under consideration, fully saturated, is something like *being morally responsible for the bad intention due to non-deterministic reasons*.

In upholding premise one of Inside-Out-style arguments, insisting that the property is intrinsic, we sacrifice either premise 2 or 3: either seemingly extrinsic properties are somehow included within the relevant intrinsic property or we are unable to determine whether we have intrinsic duplicates in worlds with differing extrinsic properties. Someone who holds onto the view that knowledge and responsibility are intrinsic will have good reason to think that extrinsic facts matter to whether someone (or their duplicate) exists in a world.

One might also reasonably reject the assumption that knowledge, or specifically *knowing there is a tree outside*, is an intrinsic property. It is undeniable that I am the knower, and so the knowledge claim is in some sense about myself and my mental state(s). But knowledge is an evaluative type or kind, and evaluative types or kinds, on the whole, depend on certain extrinsic properties or facts. Whether I have the property *knowing there is a tree outside* depends (in part) on extrinsic cooperation from the world at large.

Responsibility is also an evaluative type or kind. So, the reasons we use in the knowledge case should also apply, mutatis mutandis, to the original Inside Out argument and our Martian and fatalist parodies. Whether or not we are responsible also seems to depend on extrinsic cooperation from the world at large.

[26] The property I indicate here might not be fully saturated. If the particular property I have depends on how it was formed, we might need to get incredibly specific. A truly stringent requirement here will perhaps spell doom for our having intrinsic duplicates at all. To the extent we think the exact causal story or reasons must be included, a true duplicate will be harder to find in other possible worlds.

Bailey does not deny this general point; he does not think any responsibility property will work for the Inside Out argument. Some intrinsic properties rely on cooperation from the external world (Bailey, p. 3). But, as I've noted, he thinks that *some* mental states, including some regarding moral responsibility, will be impervious to such external variation.

Here, comparing knowledge claims with responsibility is instructive. Some knowledge ascriptions are impervious to external variation; others are not. Some knowledge claims—those which are transworld stable—withstand evil demon scenarios. *Knowing there is a tree outside* does not have transworld stability. *Knowing that I think; I exist* does. Whether there are trees outside requires a certain kind of dependence on the external world that my thinking (and thus my existing) does not. How things *seem* to me to be similarly has transworld stability: to alter how things seem to me would be to change my mental state entirely or radically resituate something in my mental economy. These seemings—and knowledge of my own existence—are appropriately internal; these are good candidates for intrinsic properties of the sort that Bailey thinks are necessary for an Inside Out argument.

The distinction between these two sorts of knowledge ascriptions is notable for our analogy. If some knowledge is transworld stable and some is not, we must carefully specify what sort of knowledge we are dealing with before we attempt to reason from the inside out. The same can be said for claims about responsibility—if some are supposedly stable and others aren't, we must be clear in articulating the difference, especially if our aim is to convince an agnostic audience.

As it happens, there is an easy way for us to identify which cases of knowledge are transworld stable and which ones aren't. In the case of my *knowing there is a tree outside*, the support is not entirely self-supplied on the basis of reflection. In the case of my *knowing that I think*, the rational support *is* entirely self-supplied by my reflective activity; I cannot be mistaken. My reason for believing—and my knowing—is given in the very activity of my reflection.

Unfortunately, this stability is unavailable when considering intentions and responsibility. There is no candidate blameworthy intention analogous to *I know that I exist*, or *I know that it seems to me that there is a tree outside*. Intentions themselves might self-supply their own existence via rational consideration, and thus are candidates for transworld stability.[27] We have reason to think that intrinsic duplicates will share all intentions. But no such self-validating reasons are forthcoming when considering responsibility: thinking one is responsible does not guarantee that this is so. Internal assignations of responsibility do not admit of certainty about the correctness of the judgment. I can believe myself not responsible for an intention when I, in fact, am. I can also—confidently—hold myself responsible when I am not. There is no clear candidate intention for which I'm obviously responsible in every instance of my having the intention. Any intention which could play this role appears to have fineness of grain that is inaccessible to me; different intentional states will appear and feel to me the same from the inside.

This lack of certainty does not entail that I'm generally unreliable in my ability to make internal assessments of responsibility, just like global skeptical scenarios do not necessarily entail that my ordinary judgments about trees or hands are unreliable. Rather, I'm not infallible with respect to these judgments, and that is enough to make it so I cannot locate an intention for which I am

[27] Though we can be mistaken about the nature of our intentions; they can be more opaque than we'd desire. Here I assume that the nature of at least some of our intentions are self-evident on reflection and provide the certainty necessary for reliable transworld stability.

responsible that has transworld stability. *Being morally responsible for a bad intention* seems more akin, even internally, to knowledge of trees rather than knowledge of my own existence. Without infallibility, I lack the ability to reliably reason across wildly varying modal contexts; I cannot reason from the inside out in a way which does not beg the question.

6 | CONCLUSION

Notably, I have refuted the Inside Out argument utilizing Bailey's preferred methodology. Bailey wants to avoid obscure theorizing if possible and reason from judgments about cases to more abstract theorizing (see p. 6). My reflection on cases like Martian manipulation is driven not by recondite theories of moral responsibility, but by commonly-held intuitions (or perhaps Moorean judgments) about responsibility—something as close to common sense (and common consensus) as we can find in the literature.

In resisting responsibility explosion, we undermine the Inside Out argument. Thus, we must bite the bullet of Martian world responsibility—one that many ardent compatibilists will reject—or abandon the argument.

It is difficult to appropriately isolate and consider internal or intrinsic properties or states, especially in relation to issues of dependence or necessity. A good (transworld-stable) duplicate is hard to find. This is not to say that we cannot competently engage in such reasoning, but we must be cautious. It does not seem, sitting in an armchair and without appropriate scientific training, that changing the speed of electron spin or tweaking the weak force ever so slightly would affect what's going on in my head or whether I exist at all. And yet, cosmologists teach us that such things do affect my existence and mental states. So, too, one might reasonably think that there is a significant difference in what intrinsic properties are possible in a world with nearly deterministic, as opposed to deterministic, laws.[28] It is not obvious that one can "turn determinism on and off" (Bailey, p. 5) without internal or intrinsic differences. Dependent beings are not wholly unmoved movers, whose intrinsic duplicates we may treat as unyielding or unvarying in highly different circumstances.

We must be especially cautious when reasoning "from the inside". Confidence in modal judgments, especially when considering changes in the laws, is hard to come by (see van Inwagen, 1998). I can be confident in my being morally responsible for a bad intention or the fact that I have hands. However, this does not grant that I'm able to be confident that I (or my duplicates) have moral responsibility or hands in other modal contexts. My confidence, even in Moorean judgment, does not grant transworld stability. When isolating my considerations to just my internal states, I encounter the problem that, in certain problem cases, things would look the same to me from the inside regardless of whether the relevant conditions or properties were present. To determine whether or not these conditions or properties are stable across worlds, we

[28] There is a general concern about how minor variations in physical laws could make all the difference regarding moral responsibility. Compare a nearly deterministic world, in which the laws are 99.99999% deterministic, and a deterministic one. How could a probability change of .00001% make such a difference? (See Fischer's contributions in Fischer et al. (2007) for this concern.) Setting aside general responses to Sorites series, one plausible response is that whether one is responsible is an "on or off" matter but responsibility comes in degrees. (See Nelkin (2016) for a defense of this view.) In a world with laws that are 99.99999% deterministic, then, it's plausible to think that beings in this world can be morally responsible, but they aren't responsible to a very high degree. Whether it's possible to have an intrinsic duplicate in a world with radically different probabilistic laws remains a fruitful direction for future work.

must look to things beyond the "inside out" perspective. And that is just what the incompatibilist says we must do.

ACKNOWLEDGMENTS

Thanks to Nathan Ballantyne and Alicia Finch for helpful discussion and incredibly generous written comments. I'm also grateful to an audience at the 2022 Pacific APA in Vancouver, and to Kevin Timpe for being my philosophical "phone-a-friend" for some references in the contemporary moral responsibility literature.

FUNDING INFORMATION

Funding was provided by Fordham University.

ORCID

Amy Seymour https://orcid.org/0000-0001-9693-6898

REFERENCES

Bailey, A. M. (2021). Compatibilism from the inside out. *Analytic Philosophy*, 1–10. https://doi.org/10.1111/phib.12227

Clarke, R. K., McKenna, M., & Smith, A. M. (Eds.). (2015). *The nature of moral responsibility: New essays*. Oxford University Press.

Fara, M. (2005). Dispositions and habituals. *Noûs*, *39*, 43–82.

Fischer, J. M., Kane, R., Pereboom, D., & Vargas, M. (2007). *Four views on free will*. Blackwell Publishing.

Fischer, J. M. (2004). Responsibility and manipulation. *The Journal of Ethics*, *8*(2), 145–177.

Johnston, M. (2016a). The personite problem: Should practical reason be tabled? *Noûs*, *50*(4), 617–644.

Johnston, M. (2016b). Personites, maximality and ontological trash. *Philosophical Perspectives*, *30*, 198–228.

Johnston, M. (1992). How to speak of the colors. *Philosophical Studies*, *68*, 221–263.

Kapitan, T. (1986). Deliberation and the presumption of open alternatives. *The Philosophical Quarterly*, *36*(14), 230–251.

Kaiserman, A. (2019). Stage theory and the personite problem. *Analysis*, *79*(2), 215–222.

Lewis, D. (1983). Extrinsic properties. *Philosophical Studies*, *44*, 197–200.

Lewis, D. (1997). Finkish dispositions. *The Philosophical Quarterly*, *47*, 143–158.

Markosian, N. (1999). A compatibilist theory of agent causation. *Pacific Philosophical Quarterly*, *80*, 257–277.

Mickelson, K. (2017). The manipulation argument. In K. Timpe, M. Giffith, & N. Levy (Eds.), *The Routledge companion to free will* (pp. 166–178). Routledge.

Nelkin, D. (2016). Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs*, *50*(2), 356–378.

Pereboom, D. (2001). *Living without free will*. Cambridge University Press.

Smart, J. J. C. (1961). Free-will, praise, and blame. *Mind*, *70*(297), 291–306.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 187–211.

Van Inwagen, P. (2008). What does an omniscient being know about the future? *Oxford Studies in Philosophy of Religion*, *1*, 216–230.

Van Inwagen, P. (1998). Modal Epistemology. *Philosophical Studies*, *92*, 67–84.

Van Inwagen, P. (1983). *An essay on free will*. Oxford University Press.

Zimmerman, M. J. (2015). Varieties of moral responsibility. In R. K. Clarke, M. McKenna, & A. M. Smith (Eds.), *The nature of moral responsibility: New essays* (pp. 45–64). Oxford University Press.

How to cite this article: Seymour, A. (2024). Against the inside out argument. *Analytic Philosophy*, *65*, 187–202. https://doi.org/10.1111/phib.12275