Philos Stud (2022) 179:2869–2890
https://doi.org/10.1007/s11098-022-01805-6

Moral worth, right reasons and counterfactual motives

Laura Fearnley[1]

Accepted: 24 January 2022 / Published online: 7 June 2022

Abstract This paper explores the question of what makes an action morally worthy. I start with a popular theory of moral worth which roughly states that a right action is morally praiseworthy if and only if it is performed in response to the reasons which make the action right. While I think the account provides promising foundations for determining praiseworthiness, I argue that the view lacks the resources to adequately satisfy important desiderata associated with theories of moral worth. Firstly, the view does not adequately capture the degree to which an action has moral worth, and secondly, the view does not identify if right actions produced from overdetermined motives have moral worth. However, all is not lost; I also argue that the account can satisfy the desiderata when it attends to the agent's counterfactual motives in addition to their actual motives. By considering counterfactual motives, we can measure the robustness of the actual praiseworthy motive, and attending to motivational robustness allows the new proposal to fully satisfy the two desiderata. At the end of this paper, I respond to some criticisms typically brought against a counterfactual view of moral worth.

Keywords Moral worth · Praiseworthiness · Overdetermination · Counterfactual motives · Right reasons · Normal worlds

1 Introduction

According to a popular approach to moral worth, a right action is worthy of praise if and only if the agent performed it in response to the relevant moral reasons, that is, the reasons making it right. Call this the Right Reasons Thesis (RRT). The central

✉ Laura Fearnley
laura.fearnley@glasgow.ac.uk

[1] School of Humanities, University of Glasgow, Glasgow, UK

idea behind this doctrine is that moral worth is not about doing something right because it is right, rather it is about doing something right for the reasons which make it right. This paper has two primary ambitions. The first is to show that RRT is not as successful as contemporary discussions suggest. This is because the view fails to adequately satisfy two important desiderata associated with theories of moral worth:

(1) DEGREES: A theory of moral worth ought to successfully identify the extent to which an action is praiseworthy.
(2) OVERDETERMINATION: A theory of moral worth ought to identify if right actions produced from overdetermined motives have moral worth.

The second ambition of this paper is to demonstrate that RRT can satisfy the desiderata when the theory is supplemented with a counterfactual framework. Supplementing RRT with a counterfactual framework entails that when assessing an action's moral worth, we not only consider whether the agent was motivated by the right reasons in the actual world, but also whether she is responsive to moral reasons in other possible worlds. In Sect. 5, I argue that the possible worlds which are relevant to moral worth appraisals are not those which are nearby, but instead those which are comparatively normal. By aggregating the number of normal worlds the agent would act well in we can determine how strongly she is motivated by the right-making reasons; the more worlds the agent acts well in the stronger her responsiveness to the right-making reasons. I argue that it is in virtue of attending to the agent's motivational strength that the proposal is able to satisfy the above desiderata. Let us call RRT combined with a counterfactual framework the Counterfactual Right Reasons Thesis (CRRT).

In the next Section, I introduce RRT and CRRT in more detail. In Sect. 3, I outline well-known extensions to RRT which aim to capture degrees of moral worth; I argue that these extensions generate implausible conclusions. Following this, I show that CRRT generates more intuitive conclusions about degrees of moral worth, and hence, better satisfies the first desideratum. In Sect. 4, I argue that RRT problematically implies that all motivationally overdetermined actions have moral worth. I then demonstrate that CRRT is committed to a different nonproblematic claim which better satisfies the second desideratum. Finally, in Sect. 5, I respond to putative objections to CRRT by delineating the account in further detail.

To clarify, my aim here is not to defend RRT, rather my aim is to argue that if you're already an advocate of RRT, then you have strong reasons to adopt CRRT. Not only does an appeal to counterfactuals provide a more successful theory in virtue of better satisfying the desiderata, it does so in a way that is uniquely unified, intuitive and otherwise theoretically nonproblematic.

2 The right reasons thesis and the counterfactual right reasons thesis

Prominent defenders of RRT include Nomy Arpaly and Julia Markovits. Arpaly proposes that:

> [F]or an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons, that is, the reasons making it right. (2002: 226)

Similarly, Markovits writes:

> [M]y action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action—that is, if and only if I perform the action I morally ought to perform, for the (normative) reasons why it morally ought to be performed. (2010: 205)

Thus, an agent is praiseworthy so long as she is motivated by considerations that explain why her action is right—it relieves suffering, respects personhood, increases welfare, and so on. Given that RRT finds value in being motivated by the reasons that make something right, the view is often presented as a rival to Kantian accounts which, by contrast, find value in being motivated by rightness per se. Arpaly and Markovits object to Kantian accounts on the grounds that they are unreasonably restrictive. To illustrate their complaint, consider the now familiar case of Mark Twain's Huckleberry Finn. Huck regards slavery as a legitimate form of ownership, he consequently feels tremendous pangs of guilt when he lies to the slave catchers about the whereabouts of Jim, a runaway slave, thereby securing Jim's freedom. In doing what he believes to be the wrong thing, Huck is not motivated by the rightness of his action, still, it seems like Huck is praiseworthy, and further, his praiseworthiness can be explained by the fact that his helping Jim is driven by a response to the relevant moral reasons—a recognition of Jim's personhood.

RRT has attracted many contemporary sponsors.[1] I suspect that a large part of the account's appeal is its ability to accommodate for the *non-accidentality constraint;* the highly intuitive thought that morally worthy actions are non-accidentally right.[2] Non-accidentality is a central feature of praiseworthy actions recognised by Kant:

> For, in the case of what is to be morally good it is not enough that it *conform* with the moral law but it must also be done *for the sake of the law*; without this, that conformity is only very contingent and precarious, since a ground that is not moral will indeed now and then produce actions in conformity with the law, but it will also often produce actions contrary to law. (1997: 4:390)

Kant rightly notes that morally worthy actions must be issued from a motive that is sufficiently grounded in the right sorts of considerations, otherwise the motive would not be reliable at generating morally right actions. RRT is said to satisfy the constraint because it demands that one ought to perform an action in response to the reasons for which it ought to be performed, thus ensuring a tight connection between motives and morality. The importance of satisfying the non-accidentality constraint cannot be understated; theories are often evaluated in terms of whether they can successfully accommodate for the idea, for if they bestow moral credit upon a wide range of lucky cases we have decisive grounds to reject the view. For

[1] For example, Amy Massoud (2016), Errol Lord (2017) and Daniel J Miller (2018).

[2] I borrow the term 'non-accidentality constraint' from Jessica Isserow (2019).

instance, if a view were to ascribe praiseworthiness to a person who saves a life only in the hope that their name will be featured in the local paper, then the view ought to be rejected. It would be a mistake to attribute praise to someone who saves a life only because doing so happens to coincide with their self-interested desires.

A second reason for RRT's popularity is entailed by the fact that moral knowledge is not required for moral worth. It doesn't matter if I know the right reasons or if I believe that I am acting for these reasons, all that matters is that *I do in fact* act for these reasons. As a result, people like Huckleberry Finn, who do something morally right whilst believing themselves to be acting wrongly, deserve moral credit (RRT's verdict on the Huck case is often considered a significant virtue of the account).[3]

Despite its strengths, we shall see that RRT lacks the resources to fulfil important desiderata associated with theories of moral worth. Before turning to these, however, I will introduce CRRT, though the introduction will be brief because it will become clear what a fully-fledged account looks like as we go along. For now, I'll say that CRRT maintains RRT insomuch as praise requires doing right for the right reasons, but unlike RRT, CRRT demands that the agent not only be responsive to such considerations in the actual world but that they continue to be responsive in a range of counterfactual scenarios. More formally:

> *Praiseworthiness:* For an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons in the actual world, *and* for it to be the case that she would do the right thing for the relevant moral reasons in a range of relevantly similar counterfactual scenarios.

Precisely how many counterfactuals make up a range and what is meant by relevantly similar will be outlined in Sect. 5. For now, it's enough to say that CRRT turns moral worth simpliciter into a threshold concept whereby agents must clear a threshold by possessing a modest degree of motivational robustness which guarantees they will act well in a handful of similar circumstances.

By demanding even a modest amount of motivational robustness, CRRT offers an immediate advantage over RRT—it's more effective at satisfying non-accidentality. As noted, RRT requires a tight connection between motives and morality, and hence, goes some way to securing the constraint. Still, in requiring that one's motive be robust enough to elicit the action in a range of different scenarios, CRRT demands an even tighter connection between motives and morality.[4] This virtue might not provide a decisive reason to accept CRRT, but given the importance of the constraint, it does offer an initial motivation for

[3] For discussions regarding the infamous Huck Finn case see, for example, Bennett (1974), Montmarquet (2012) and Sliwa (2016).

[4] Isserow notes that a counterfactual view of moral worth is more successful than accounts like RRT in capturing the type of non-accidentality constraint I sketch out here because "non-accidentality seems to require some measure of counterfactual robustness" (2019: 256). Isserow also categorises this type of counterfactual threshold view as a "strong dispositional view" (2019: 254). I do not adopt this label because CRRT does not aim to measure one's overall disposition to act well, rather it aims to measure the robustness of a particular motive.

preferring the view. Having now outlined both accounts, I will turn to discuss the desiderata, I begin with DEGREES.

3 DEGREES

Sometimes two seemingly morally right actions possess different amounts of moral worth. For example, we may say that Jane deserves more praise for baking her friend a birthday cake than John does for baking his friend a birthday cake on the grounds that Jane is under emotional strain having recently suffered a bereavement. RRT, as it stands, does not discriminate between John and Jane because the view treats all praiseworthy actions as having equal moral worth, yet a comprehensive theory will go further in identifying the *extent to which* an action has moral worth.

In this Section, I look at two different ways in which RRT has been extended to capture degrees of moral worth—the first defended by Markovits, the second by Arpaly and Timothy Schroeder. I intend to show that both of these proposals fail to fully deliver on their promise of satisfying DEGREES. As I see it, Markovits's proposal falls short on two counts; firstly, it doesn't capture the full spectrum of degrees of moral worth, and secondly, it doesn't seem to establish an action's degree of praise*worthiness*. Arpaly and Schroeder's proposal, on the other hand, commits us to unintuitive conclusions about degrees of moral worth. Having raised these objections, I go on to offer CRRT's alternative solution.

3.1 DEGREES and RRT

In order to satisfy DEGREES, Markovits combines RRT with an appraiser-relative approach to moral worth. How much praise an action is owed depends upon how we, as appraisers, would have acted had we been in the agent's shoes. In regards to maximally praiseworthy actions, Markovits states:

> A heroic action is a right action (of some moral significance) that most of us, judging the action, would not have had the moral strength to perform, had we been in the hero's place. (2012: 297)

Hence, the extent to which an action deserves moral worth is relative to a community of appraisers—the more unusual it would be for that action to occur in one's moral community the more admiration it deserves. But, of course, there can be disagreements amongst communities. I might judge the fireman to be a hero because I would not have risked my life had I been in his shoes, still his colleagues could reject the compliment; 'he's only doing his job' they might protest. Markovits suggests that when this type of disagreement occurs that both of these assessments could be right, and therefore, one can be appropriately described as both a hero and not a hero at the same time (2012: 297).

Markovits presents an interesting extension to RRT. However, I worry that it doesn't completely capture the full spectrum of praiseworthiness. The account offers the resources to identify maximally praiseworthy action—actions that most of us would not have performed ourselves. But what of behaviour that falls short of this extreme? A friend of the account could delineate the proposal further to capture these more ordinary actions. For instance, one could say that if a heroic action is one that most people judging the action would not have the moral strength to perform themselves, perhaps a considerably praiseworthy action which falls short of heroism is one that nearly most of the people judging would not have the moral strength to perform themselves. By way of example, if 90% of people consider themselves unwilling to replicate the behaviour of the agent, then the act warrants maximal moral admiration, whereas, if only 70% judge themselves unwilling then the action is certainly praiseworthy but not heroic. This method aggregates moral appraisals to find a kind of mean which then corresponds to the degree of moral worth.

Although initially plausible, it's not obvious to me that an appraiser-relative view is compatible with aggregating judgements in this way. For recall, that if communities disagree about whether an action is heroic or not heroic, then the action itself might appropriately be described as heroic and not heroic. This implies that moral appraisals cannot be combined to generate new moral appraisals; we cannot take a heroic judgement, add it to a non-heroic judgement to get 'almost heroic'. In other words, moral appraisals cannot be aggregated with a view to working out the mean. If this is the case, then the extension seems to fall short of fully satisfying the desideratum, for it only explains when an action deserves maximal credit, leaving a large swathe of more commonplace behaviour unaccounted for.

My second, and perhaps deeper worry, concerns whether the appraiser-relative approach captures praise*worthiness*. On Markovits's view an action is more praiseworthy relative to appraisers, this means that degrees of moral worth depends upon how appraisers stand in relation to the action; namely, whether they would have had the moral strength to perform that action. However, moral worth is typically understood to be a feature of an action that goes above and beyond standing relations. To illustrate, suppose that Wonder Woman performs a dangerous rescue to save a group of children. We don't think her action is made less praiseworthy by the fact that Superman would have performed the same dangerous rescue had he been in Wonder Woman's place. We might think it would be *inappropriate* for Superman to *praise* Wonder Woman by, for example, applauding her after witnessing the rescue because this gesture comes across as condescending or disingenuous given how Superman stands in relation to the action. But questions about whether it's appropriate to praise someone are importantly different from questions about whether an action is genuinely worthy of praise.[5] To my mind, the extent to which an action deserves praise is independent of how appraisers stand in relation to the action, and if this is the correct way to think about praiseworthiness, then perhaps the agent-relative approach is the wrong way to determine it.

[5] In recent years, there has been a flurry of work on so-called standing to blame which involves identifying facts about the blamer that are relevant to whether an instance of blame is appropriate. There appears to be much less said on standing to praise (with the exception of Kasper Lippert-Rasmussen (2021)). Nonetheless, it's reasonable to suppose that, like the relationship between blame and blameworthiness, there will be instances where someone is praiseworthy but praising them is inappropriate given the praiser's standing.

However, some might remain unmoved by this worry. One could continue to hold onto the thought that degrees of moral worth is a relativised feature of an action and that the standing relation accurately tracks this feature. Even so, I think taking up this line of argument would be difficult if one were an advocate of RRT. This is because RRT supplies conditions for moral worth simpliciter which do not relativise the phenomenon; whether an action is praiseworthy or not praiseworthy is determined independently of subjective judgements and beliefs at particular times and places. Hence, if one wanted to advocate for both RRT and an appraiser-relative approach, some story has to be told as to why degrees of moral worth is a relativised feature of action yet moral worth simpliciter is not. In the absence of such an explanation, I think we have reasons to be cautious about adopting the view.

So much for the appraiser-relative extension to RRT, let's now consider a different approach developed by Arpaly and Schroeder. Arpaly and Schroeder take the degree of praiseworthiness to depend on how strongly the action manifests an intrinsic desire for the right-making features. Since strength of desire seems like the type of thing that can be scalar, it easily explains how praiseworthiness can come in degrees: one whose good action manifests a stronger desire for the good is more praiseworthy than one who manifests a weaker desire.

This account may strike you as similar to the one I aim to develop here. After all, in the introduction I stated that CRRT will meet the desideratum by measuring how strongly one is motivated by the right sorts of reasons. However, there's an important feature of Arpaly and Schroeder's view that differentiates it from CRRT—it does not attend to counterfactual motives but only the strength of desire which is *actually manifested* in action. To illustrate the idea of actual desire manifestation, Arpaly and Schroeder ask us to imagine two agents who kindly give a lost motorist directions; the first agent is a moral saint with bottomless good will, whilst the second has a quite average amount of good will. For Arpaly and Schroeder, "an opportunity to assist a lost motorist is not typically an occasion for a full display of a powerful commitment to morality. Hence, the strength of the desire for the right or good that is actually manifested in the two cases we imagined is the same" (2013: 189). Accordingly, the two agents deserve the same amount of moral credit, despite the fact the first agent generally possesses more good will than the second.

On other occasions, the interior life of the agent can present opportunities to display a powerful commitment to morality, thus making a good action more praiseworthy than it would ordinarily be. This happens, for example, when a person experiencing depression continues to do good despite undergoing great sadness. It takes a strong desire to respond to moral reasons in the grips of depression, and so assuming that this desire is manifested in her actions, "the sorrowing agent is more praiseworthy for her action than a person would be for doing the same good works without having to overcome the same psychological barriers" (2013: 189).

Whilst focusing on the strength of desire manifested in action nicely tracks our judgements in these types of examples, I think it fails to do so in other cases, namely, in cases where agents possess a strong competing self-interested desire to act otherwise. To illustrate my concern, consider the following example:

> DONATION: Two agents, Lola and Kirke, receive a £500 work bonus. Shortly after receiving the bonus, their employer reminds them of the company affiliated charity—UNICEF. Both agents decide to donate their bonuses to UNICEF and both do so for the right sorts of reasons, but Lola and Kirke experience very different internal processes before they come to this decision. Kirke feels a variety of self-interested desires to spend the money on himself, 'after all', he reasons, 'I've earnt this money through hard work, I ought to treat myself'. Kirke's desires to keep the money are strong; it takes him a few hours of painful deliberation and pacing before he overcomes his internal struggle and is able to donate. Lola, on the other hand, feels no internal resistance or temptation to spend the money on self-interested pursuits. After receiving her bonus, she swiftly gives it to UNICEF.[6]

In DONATION, both agents perform the same morally desirable action in response to the right sorts of considerations, but we can suppose the strength of concern manifested in their respective actions are different. To put it somewhat artificially, suppose that Kirke's desire for the right-making reasons has a strength of 50 and his desire to keep the money has a strength of 49, while Lola manifests a desire of 40 for the right-making reasons and has no self-interested competing desire. Given that Kirke's donation manifests a stronger desire for the good, an account, like Arpaly and Schroeder's, that posits strength of desire manifestation as a criterion for degrees of moral worth would determine that Kirke deserves *more* moral credit than Lola.

I think that this conclusion is too quick. For one thing, it seems obvious that Kirke should not receive special admiration just because he eventually managed to resist temptation. To argue otherwise would be to penalise Lola for lacking such temptations in the first place. Moreover, praising Kirke more than Lola would be especially dubious if we think desires are the types of things we can have agency over. If we have the power to regulate and reform our desires, then a person who finds it difficult to do well because they have failed to appropriately govern their self-interested desires, should not, other things being equal, be given more moral credit compared to a person who finds doing well easy as a result of the fact they've effectively regulated their desire profile.

A defender of the desire manifestation view might be tempted to debunk my intuition that Kirke is not more praiseworthy than Lola by appealing to judgements about Lola's character. The response might go like this: the fact that Lola donates her money with ease provides evidence to suggest that she's a good *person* and this thought distorts evaluations of how much praise she deserves for her *action*. Namely, it leads us to think that she deserves more praise than she actually does. This would be problematic because we would be assigning value not only to the good motive but also to the feature which makes it easy for her to act on this

[6] This example is one adapted from Kelly Sorensen (2010), and like Sorensen, we should imagine that Lola and Kirke share similar economic circumstances. £500 is not a trivial amount of money for them, but equally, forgoing the bonus will not deprive them of any necessities.

motive—her character. Hence, we would be conflating moral worth with moral virtue.[7]

Although it's important to bear in mind these distinctions, I don't think this explanation debunks the targeted intuition because the intuition does not rest on a mistaken conflation between types of moral appraisals. We can see this by filling in the details of the case in a way that makes it clear that Lola's score on the character dimension of appraisal does not unduly inflate her score on the moral worth dimension of appraisal. To do this we simply stipulate that Lola, in fact, has a subpar moral character and that when she donates her bonus she acts out of character. In this story, Lola is occasionally generous, kind, honest, etc., but for the most part, she experiences desires that push her towards the morally neutral, and occasionally the morally bad. She certainly does not typically display the kind of generosity required to donate £500 to UNICEF. Despite this, on the day her bonus arrives, she forms an uncharacteristic urge to relieve the suffering of those less fortunate, thereafter she donates the money with ease. Although I've now specified that Lola's character is somewhat substandard, I take it that our judgements about how much moral credit she deserves remain the same; she's just as praiseworthy for giving away her bonus irrespective of whether she's a virtuous person or not. With this line of argument dispelled, we're back to the thought that praising Kirke more than Lola would be an error, thus I think we have reason to believe that the strength of desire manifested in action does not track the degree of an action's moral worth.

3.2 DEGREES and CRRT

I think an alternative solution to satisfying DEGREES can be found by taking a step back to consider the fundamental desideratum on moral worth—the non-accidentality constraint. Recall that moral worth simpliciter depends upon whether an action was brought about through luck. If it's accidentally right, then it's not a candidate for moral worth. It's reasonable to suppose then, that the degree of moral worth depends on the degree to which the action depended on luck. One way to capture the degree of luck involved in action is by looking at how the agent would have acted in various counterfactual scenarios. If the agent would continue to do well in a broad range of counterfactual scenarios, then certain circumstances were not needed to bridge the gap between motivation and rightness. For such an agent, her praiseworthy motive plays a leading role in generating action, we can be sure that it's her motive and not the environment which is worthy of credit. Whereas, if an agent fails to do well in lots of alternative scenarios, then certain circumstances were needed to forge the connection between motivation and rightness in the actual world, thus, her action is dependent more on luck.

To illustrate this thought, suppose that Kirke's good motive is moderately robust; he's able to overcome his self-interested desires and donate to UNICEF in many relevantly similar scenarios. In his case, Kirke's actually doing well is obviously no

[7] The objection that a counterfactual account of moral worth tracks moral character instead of moral praise has been advanced by many including Herman (1981), Markovits (2010) and Isserow (2019).

accident. If, by contrast, his motive was precarious enough such that he would fail to donate in slightly different scenarios, say, in ones where he's hungry, irritable or he forgets his online banking login details, then his action warrants little praiseworthiness. His actual action seems almost accidental, creditable to his remembering his banking details and his employer's prompt more so than his desire to relive the suffering of others.

I propose then, that DEGREES is solved by attending to how strongly the agent was motivated to respond to the relevant moral reasons, where strength is cashed out in terms of counterfactual robustness. Simply put, the more counterfactual situations one would continue to perform the same desirable action in, the more robust the motive, and thus, the more praise one deserves. Conversely, the fewer counterfactual situations one would continue to perform the same desirable action in, the more fragile the motive, and thus, the less praise one deserves. In short, the amount of moral worth awarded is proportional to the robustness of one's motive, and how robust an agent's motive is acts as a proxy for something more important—to what extent the action is a product of accidentality.

With this new condition in place, let us take stock of what has been said about CRRT thus far. In Sect. 2, I stated that CRRT fashioned moral worth into a threshold concept—one must respond to the right sorts of reasons not only in the actual world but also in a range of possible worlds to gain moral worth simpliciter. Combining this with what I've said about DEGREES, it follows that once an agent has met this threshold, we can move to ask how many other worlds she would do well in, the more of these other worlds she would do well in the more praise she deserves.

4 OVERDETERMINATION

I use the term overdetermination for cases in which one has two or more independent motives for doing the right thing and would have acted rightly from any one of those motives even in the absence of the others. Had one motive not been present the agent would have acted anyway. The category of overdetermined actions which presents difficulties for a theory of moral worth are those in which one of the motives is praiseworthy and the other is not praiseworthy. For advocates of RRT, the specific worry will arise when an agent does the right thing for the reasons which make it right whilst also being moved by reasons which do not make it right. To illustrate, imagine a politician volunteers to help at a food bank, and she has two motives for doing so:

M1: It is in her career interest to be seen volunteering.
M2: She desires to relieve the suffering of those less fortunate.

Supposing that M2 constitutes doing something right in response to the right reasons, are we to say that on this occasion her action was one done in response to the right reasons, and thus, had moral worth?

In this Section, I evaluate some answers to this question. I demonstrate that RRT's answer is problematic because it risks violating the non-accidentality constraint. Hence, by meeting one desideratum (OVERDETERMINATION), RRT violates a different and perhaps more fundamental desideratum. I next show that CRRT maintains a different claim about overdetermined actions, and unlike the claim RRT is committed to, this claim captures OVERDETERMINATION without violating the non-accidentality constraint, therefore providing a more successful solution.

4.1 OVERDETERMINATION and RRT

As it stands, RRT is committed to something like the following:

All: All motivationally overdetermined actions have moral worth when at least one of the motives was a response to the right sorts of reasons.

What makes a right action morally praiseworthy according to RRT is the fact that the agent responded to the reasons which make the action right, nothing in the account rules out actions as praiseworthy in virtue of the person having additional motives for doing what they do. Consequently, the view entails that all motivationally overdetermined actions are worthy of praise, on the condition that at least one of the motives was a praiseworthy one. Turning to the politician case, RRT would maintain that the politician is praiseworthy for volunteering because at least one of her motives, M2, is a response to the right sorts of reasons.

Before moving on, it should be noted that as far as I'm aware prominent defenders of RRT have fallen silent on the question of overdetermination except for Markovits who writes in a footnote: "if there are cases of motivational overdetermination, it may be okay to have some nonmoral motivations for doing the right thing, so long as we're also fully motivated by the actual normative reasons justifying the act" (2010: 238, fn. 66). This brief remark gives little guidance on the question at hand other than to indicate that the theory is amendable to the idea that overdetermined actions may be praiseworthy provided that the agent is fully motivated by the relevant sorts of considerations, though it's unclear what being fully motivated entails. In any case, in the absence of any detailed discussion, I think it's fair to categorise the account as endorsing *All*.[8]

In her influential paper, Barbara Herman points out that endorsing *All* is problematic because we would end up praising some actions which are only accidentally right. Here is what she says on the matter:

> As circumstances change, we may expect the actions the two motives require to be different and, at times, incompatible. Then […] an agent might not have a moral motive capable of producing a required action "by itself" if his presently cooperating nonmoral motives were, instead, in conflict with the moral motive. That is, an agent […] could, in different circumstances, act contrary to duty, from the same configuration of moral and nonmoral motives that in felicitous circumstances led him to act morally. (1981: 367)

[8] *All* also seems to be in the spirit of what Markovits proposes in her footnote.

Here Herman argues that when circumstances change, we may expect the two motives that were hitherto compatible to become antagonistic, pulling the agent towards different ends. During such conflict the agent may feel the pull of the non-praiseworthy motive more than the praiseworthy one leading them to act contrary to duty. Attending to the possibility that the agent would *not* act well in the altered circumstances introduces the suspicion that the original configuration of motives produced right action only accidentally. The conditions of cooperation between the two motives which led to right action in the actual situation depended upon the fortuitous alignment of favourable circumstances. These actions are therefore more a function of the accidental circumstances and less a function of the praiseworthy motive. Hence, to praise such performances is to praise only accidentally right actions.[9]

To clarify the problem, consider our politician again. The politician is motivated to volunteer from a praiseworthy motive, M2, and a motive of self-interest, M1. Is she praiseworthy? Possibly not. The fact that she volunteers in this world from these motives is compatible with the thought that if these two motives were no longer pushing her towards the same end she would fail to volunteer. It's easy to conceive of scenarios that make these motives combative rather than cooperative. Suppose that the politician was instrumental in enacting punishing welfare reforms which caused a dramatic increase in food bank usage. On the day the politician is scheduled to volunteer, she learns that the press no longer intend to publish a flattering story about her good deed, instead they intend to run a story accusing her of being a hypocrite for volunteering at a food bank in light of the fact that her policies made them necessary. Now if the politician would fail to volunteer in a world where she would receive negative publicity for doing so, then it reveals that her doing well in the actual world depended upon her two motives uniting in the way they did, and the reason they unite in the way they did is due to certain contingent circumstances obtaining. When the politician acts rightly in the actual world then, it is not because of a robust praiseworthy motive, but because accidental circumstances which happened to be favourable in producing right action obtained at the time. Thus, when she acts rightly, she does so somewhat accidentally.[10]

[9] Herman deploys her insights about overdetermination to argue that an action has moral worth when the primary motive for the action is the motive of duty. I set aside the wider context of her project to focus on her claim that overdetermined actions can be accidentality right.

[10] One might wonder why I am using Herman's test to check for accidentality in overdetermined actions. Herman's test entails that we consider a scenario where the non-praiseworthy motive is combative rather than cooperative. But, as Benjamin Ferguson (2012) points out, another way to test for accidentality is by considering a scenario in which the non-praiseworthy motive is simply absent. Applying this thought to the politician case, one could ask: why consider a scenario in which volunteering would be damaging for the politician's career interest, as opposed to a scenario in which volunteering is neutral with regards to her career interest? (For example, why not imagine a world where the press do not cover the story at all). In response, I note that in overdetermination cases, the agent treats the non-praiseworthy motive as a relevant reason (and not a mere cause) for or against acting in the actual situation. In the actual world the politician takes the fact that volunteering will improve her career as a reason to volunteer. It therefore seems entirely legitimate to consider alternative cases where the non-praiseworthy motive continues to be *present* and not merely cases where the non-praiseworthy motive is absent. Another way of putting it is if

In light of this, we have strong reasons to reject *All*. Whilst this claim does allow RRT to satisfy OVERDETERMINATION, it does so at the cost of violating a more fundamental desideratum on moral worth.

Once *All* is dismissed it might be tempting to consider an alternative that says no motivationally overdetermined actions are compatible with praise. Call this claim *None*. To hold this option is to argue that whenever a non-praiseworthy motive cooperates with a praiseworthy motive to bring about action, the mere presence of the non-praiseworthy motive renders that action devoid of moral worth. Kant has often (although perhaps uncharitably) been viewed as an advocate of this view.[11] He seemingly claims that a dutiful act can have moral worth only if it is done from the motive of duty alone i.e., is not overdetermined.[12] The view has thus been heavily criticised for the apparent consequence that it judges a resentfully performed dutiful act as morally preferable to a similar act done with enjoyment. If I help a friend move house because I promised and because I enjoy helping, my good deed warrants no moral credit according to this Kantian view, since my act is not done solely from duty but also from enjoyment. So whilst *None* does fulfil the desideratum, and plausibly does so without violating the non-accidentality constraint, it comes at the expense of our widely held intuitions about moral worth.[13] For this reason, we ought not place an indiscriminatory ban on actions as candidates for praise in virtue of the fact that a praiseworthy and non-praiseworthy motive were each individually sufficient to bring about its performance. Hence, we should also rule out *None* in the search for some better alternative.

The final option available to us I call *Some*: some motivationally overdetermined actions are compatible with moral worth. In particular, the set of actions that are compatible with praise are the ones that do not violate the non-accidentality constraint. In what remains, I will explain how CRRT accurately captures *Some*.

4.2 OVERDETERMINATION and CRRT

Let's peddle back. CRRT says that a right action has moral worth if it's performed in response to the right reasons not only in the actual world but also in a range of possible worlds, and to capture degrees we look to see how many additional worlds

Footnote 10 continued
the agent in overdetermined cases takes their non-praiseworthy motive as supplying relevant action-guiding reasons, then our test for accidentality ought to include such reasons.

[11] More recently, Philip Stratton-Lake has endorsed *None*. Following Herman, Stratton-Lake argues that we cannot praise all motivationally overdetermined actions since doing so would risk violating the non-accidentality constraint. And further, he finds no plausible way of being able to discriminate between those sets of overdetermined actions which violate this constraint and those which do not. He thus resigns himself to the conclusion that "overdetermined acts cannot, therefore, have moral worth" (2000: 108).

[12] In the *Groundwork for the Metaphysics of Morals*, Kant famously says of the man who is so overcome by sorrow that he is no longer moved by the needs of others: "suppose that now, when no longer incited to it by any inclination, he nevertheless tears himself out of his deadly insensibility and does the action without inclination, simply from duty; then the action first has its genuine moral worth" (1997 4:398).

[13] Many commentators have sought to amend Kant's proposal in order to avoid this aspect of the theory. For an interesting discussion see Henson (1979), Herman (1981) and Benson (1987).

the agent would act well in. So how does CRRT satisfy *Some*? The overdetermined actions which are compatible with praise are simply the ones that meet the condition for moral worth simpliciter. In meeting this condition we can be certain that their actually doing well was not the product of accidental circumstances which fostered cooperation between the praiseworthy and non-praiseworthy motive—luck could not persist across modal universes in this way. If, on the other hand, the praiseworthy motive was precarious enough to the extent it could easily be overridden by non-praiseworthy motives in most similar scenarios, then the agent's actually doing well was a result of the accidental cooperation of praiseworthy and non-praiseworthy motive, therefore, they deserve no moral credit.

To clarify, consider our politician again. Recall that the problem with praising her for volunteering was the thought that she would not volunteer in a world where M1 and M2 no longer cooperated, that is, in a world where volunteering conflicts with her career interest. What does CRRT say about this case? Generally, it says that the politician is praiseworthy if her moral motive were sufficiently strong enough to see her volunteer in a range of relevantly similar scenarios, but she is not praiseworthy if her motive is sufficiently weak such that she would fail to volunteer in these scenarios. To find out if the politician's action deserves praise then, we must get precise about what counts as a relevantly similar scenario and how many of these scenarios constitutes a range. I attempt to do this in the next section.

5 Ranges and relevant counterfactuals

According to CRRT, moral worth simpliciter requires possessing a somewhat robust praiseworthy motive. I've cashed out robustness in terms of counterfactuals such that moral worth simpliciter requires acting rightly in response to the right reasons not only in the actual circumstances but also in counterfactual circumstances. The degree of moral praise is also determined by how an agent would have been motivated had things been different. But appeals to counterfactual motivations might strike you as odd insomuch as possessing or failing to possess a praiseworthy motive in some alternative scenarios doesn't seem to matter to the moral worth of the actual action. Consider the following example:

> Aisha runs a marathon for charity in the actual world, but had she fallen at the start line and broken her ankle, she would have been motivated differently and failed to compete in the race as a consequence.

If we maintain that evaluations of moral worth are sensitive to non-actual motivations, then we might conclude that whatever amount of credit Aisha deserves is mitigated by her failure to do well in the broken-ankle-world. I agree that this would be the wrong conclusion. In response, one might be tempted to reject appeals to counterfactuals altogether, but given that non-accidentality seems to require some measure of motivational robustness, I think this move is overhasty. A more amicable solution, and the one I undertake here, is to identify the counterfactuals that matter and those that do not.

For a first pass, we might think that the counterfactuals which matter are those instantiated in nearby worlds—worlds most similar to the actual world. Accordingly, moral praise simpliciter is gained by acting rightly in the actual world and a range of nearby possible worlds. If the agent continues to act rightly in nearby worlds after passing the threshold they deserve more moral praise. Privileging nearby worlds has initial plausibility because our moral worth judgements appear to be sensitive to motivational changes that occur in very similar conditions, while they don't appear to be sensitive to changes that occur in radically different conditions. For instance, we seem to care if our marathon runner would be motivated differently had she been unable to wear her favourite running top, because a failure to do well in this world shows that her good motive, while grounded in the right sorts of reasons, was not sufficiently strong. But we don't care how she would be motivated when circumstances are drastically different. We're not interested, for example, what she would do had the marathon taken place in an apocalypse because this world is too modally distant to render any important information regarding her actual action.

Despite its initial plausibility, I think that privileging nearby worlds would be a mistake. This is because nearby worlds occasionally contain circumstances that are significantly more demanding, and when they do the counterfactual test misfires. This happens when a nearby world's slight deviation from the actual circumstances leads to a confounding turn of events which, as a consequence, demands a greater personal sacrifice from the agent to do the right thing than originally expected. In such cases, one could hardly deny that someone's actual action lacks moral worth just because, had the moral stakes been significantly higher, they would be unwilling to perform that action.[14] I think this thought explains why we take a lot of counterfactuals to be irrelevant to decisions about moral worth. It explains our readiness to ignore Aisha's counterfactual motive in the broken-ankle-world. The broken-ankle-world is a relatively nearby one (not much has to change for us to get there, perhaps the strategic placement of a shoelace), but what unfolds as a consequence makes doing the right thing vastly more difficult, and thus, we take the scenario to have no bearing on the moral worth of what she actually does. The counterfactual test aims to identify if the motive is robust enough to transcend very particular circumstances, it's not intended to identify if the motive is unbreakable.

If the relevant counterfactuals are not those instantiated in nearby worlds, then which ones are relevant? Here is one suggestion. When we decide on an action's moral worth, we implicitly associate the action with a set of conditions that we take to be *normal* for its performance. The same is true of those doing the performing; action guiding-decisions are made in light of our expectations about how things would normally turn out. Aisha's decision to run a marathon for charity, for instance, is made with the reasonable expectation that she will not be required to

[14] Markovits has objected to a counterfactual account on these grounds. She states that "we should not think [an action is] *less* worthy because the agent who performs it (still for the same right-making reasons) might *not* have done so had the cost been higher" (2010: 213). However, this objection does not show that a counterfactual account of moral worth is incorrect but rather that not *all* counterfactuals are relevant to moral worth.

perform her good deed having sustained a severe injury. In light of this, when we consider how the agent would have acted had things been different, we ought to fix the normal conditions which contextualise the performance of the action. Manifestly, this means considering counterfactual scenarios that are instantiated in worlds that are comparatively normal from an actual world perspective.

Ranking worlds according to their comparative normalcy is not new, although it is not conventional either. The idea has been explored by metaphysicians like Menzies (2004), McGrath (2005), and Halpern (2016) in connection with a counterfactual analysis of causation, and by Smith (2007, 2010) in connection with epistemic justification and ceteris paribus conditionals. I take my lead from these authors in characterizing the notion of a normal world.

First, notice that the concept of normality is interestingly ambiguous. It has both a statistical and prescriptive element. To say something is normal in the statistical sense is to say that it conforms to a statistical mode. For example, in Scotland the winter months are generally rainy and overcast, so if Scotland were to have a sunny, dry winter, the country's weather would violate a statistical norm. By contrast, to say something is normal in a prescriptive sense is to say that thing follows a prescriptive rule. These rules are constituted by the way things *ought* to be or are *supposed* to be. Prescriptive norms can take many forms. Some norms are moral; for example, it's generally believed that people are supposed to keep their promises, even if no explicit laws or rules demand this behaviour. There are also norms of etiquette that establish standards of conduct in certain social contexts. Laws too can create norms that produce expectations about how people regulate their behaviour in societies. Policies enacted by institutions can be norms; for example, a company may have a dress code policy creating expectations around how employees dress for work. As McGrath (2005) points out, there are also norms of proper functioning for organisms and machines. Alarm clocks are supposed to ring at their set times and human hearts are supposed to pump blood around the body, and there's a sense in which 'supposed to' has normative force here; failure to function properly is a failure to meet a certain kind of standard. Broadly speaking then, a world can be categorised as normal to the extent that it abides by actual world statistical and prescriptive norms. A world where Scotland is rainy and overcast in winter is more normal, other things being equal, than a world where Scotland is sunny and dry in winter. Likewise, a world where people are expected to keep their promises is more normal, other things being equal, than a world where people are encouraged to break them. Abnormal worlds, by contrast, are those which deviate from what we expect to happen in both a statistical sense as well as a prescriptive sense.

Restricting the counterfactuals that matter to those in normal worlds, has an immediate advantage over privileging nearby worlds—normal worlds typically do not contain significantly more morally demanding scenarios. In normal worlds, circumstances evolve in ordinary ways, this inhibits confounding changes of events from occurring which in turn prevents drastic changes in the moral stakes. Aisha's broken-ankle-world, for instance, would not enjoy membership in the normal worlds, since breaking one's ankle immediately before running a marathon is statistically abnormal. Privileging normal worlds has another virtue; given that normal worlds share the same norms as the actual world, it appeases the intuition that the scenarios which matter to moral worth are those which are relevantly similar to the actual scenario.

In terms of moral worth then, we're looking to see if the agent would continue to perform the same morally desirable action in alternative circumstances that are comparatively normal from the perspective of the actual world. Having identified the relevant counterfactuals, we are now in a position to specify the necessary and sufficient conditions for moral worth under CRRT:

> *Praiseworthiness:* For an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons in the actual world, and for it to be the case that she would do the right thing for the relevant moral reasons in a range of normal worlds. Once this threshold is met, the more normal worlds the agent would continue to perform the right action in from a response to the relevant moral reasons, the more praiseworthy the action.

With the relevant worlds identified as normal, CRRT can evade a second criticism typically brought against counterfactual accounts. The objection states that in virtue of rendering morally worthy actions counterfactually robust, moral worth becomes too hard to achieve. This point has been pressed by Jessica Isserow and Paulina Sliwa. Isserow writes that "[i]n so far as one's account render's morally worthy actions counterfactually robust, it risks rendering praiseworthy agents far rarer than we take them to be" (2019: 258) and Sliwa states that "clearly, it is unreasonable to demand that to have moral worth the agent needs to have acted rightly no matter what. Some contingency must be compatible with moral praiseworthiness" (2016: 400). Isserow and Sliwa are right to point out that counterfactual accounts have the potential to be unreasonably demanding, but CRRT escapes this worry. In specifying that only scenarios manifested in normal worlds are relevant to moral worth, the account restricts the number of worlds quantified over thereby limiting the number of scenarios an agent is required to act well in. Precisely how many worlds ought to be included in the range of worlds quantified over for moral worth simpliciter will depend upon how many guarantee non-accidentality. A single counterfactual attempt to act rightly for the right reasons will not guarantee non-accidentality since it too may contain unduly biased circumstances, but sufficiently many attempts scattered across normal worlds will provide a guarantee because not all of these attempts can depend on particularly favourable conditions obtaining. In any case, the threshold will be a moderate one. An agent does not have to do rightly no matter what to gain moral credit.[15]

Let me return to the politician and deliver a final verdict on the case.[16] According to CRRT, to deserve credit for volunteering, the politician's praiseworthy motive

[15] One might press me here for a more exact answer regarding how many worlds the agent is required to act well in. Getting precise about how we quantify over possible worlds remains a standard problem for those working in ethics, epistemology and metaphysics who endorse modal theories, and it's not one I can solve here. But let me say that I think such putative precision might be illusory in any case; the very nature of possible world semantics renders supplying a more exact quantification a near impossible task, and those defending a possible world framework might have to merely accept this implication.

[16] Thanks to an anonymous reviewer for pressing me to return to this case (and others) in order to develop my own proposal.

must be robust enough to see her volunteer in a range of normal worlds, we can now ask whether worlds in which volunteering would be against her career interest represent normal states of affairs, and therefore, are relevant to establishing the politician's moral worth. To my mind, it's very plausible that at least some worlds will be normal. Consider, for instance, the world in which volunteering would gain the politician negative publicity. This state of affairs seems comparatively normal; it's statistically normal for journalists to publish stories ridiculing politicians (far more normal than publishing stories approving of politicians). Furthermore, in running the story the press are abiding by a prescriptive norm in the sense that we believe that the press are supposed to scrutinise the actions of our political representatives. Consequently, we ought to take the fact that she would fail to act well in this scenario as relevant to the politician's moral worth. Even so, considering one relevant counterfactual does not supply enough evidence to deliver a final verdict—we need to consider how she would act across a range of normal worlds. So to settle the case, let us suppose that the politician's praiseworthy motive is not especially robust; her concern for those less fortunate is not deep or impassioned but fleeting and feeble, as a result, she would fail to act well in many situations where volunteering did not coincide with her self-interest. And if we also suppose that a substantial amount of these scenarios will contain comparatively normal conditions, then according to CRRT the politician is not praiseworthy.

Let's see how CRRT handles a further two cases much discussed in the moral worth literature. Firstly, consider Markovits's example of the fanatical dog-lover who "performs a dangerous rescue operation to save a group of strangers at great personal risk" (2010: 210). Markovits argues that the dog-lover's actual rescue is not made less creditworthy by the fact he would have abandoned the strangers had his dog required his heroics at the same time. Let's assume with Markovits that saving the dog over the strangers would be the wrong thing to do. What does CRRT say about the case? Firstly, we have to determine whether the counterfactual Markovits invokes is a relevant one. Does the scenario in which the dog requires saving at the same time as a group of strangers represent a normal state of affairs? Plausibly, no. Not only would it be statistically unusual for a dog to need rescuing at great personal risk to its owner, it would be even more unusual that this should occur at the very same time a group of strangers also need saving. And I can't see any sense in which the scenario would be prescriptively normal. As a result, the counterfactual carries no weight in determining the moral worth of the dog-lover's actual action.[17]

[17] It might be possible to imagine a scenario where it would be normal for the dog and the strangers to need rescuing at the same time. We could suppose that the dog-lover walks his dog in a natural park and that more often than not dogs and hikers fall into the park's dangerous running waters. On this version of the story, there is a statistical norm according to which it would be normal for the dog and the strangers to need rescuing at the same time. Still, even on this version of the story, the dog-lover's abandonment of the strangers in favour of his dog does not significantly affect his actual praiseworthiness. Supposing that the dog-lover meets the conditions for moral worth simpliciter by rescuing the strangers in a range of normal worlds, the fact he fails to rescue the strangers in the handful of worlds where it's normal for the dog and the strangers to require his heroics at the same time, would not make much difference to how much praise he deserves overall.

Finally, consider Isserow's example of the devoted parents. These parents make great personal sacrifices for their children from a concern for their children's wellbeing, but they are so devoted that they would promote their children's wellbeing even when doing so is morally wrong. For example, they may refuse to let their children experience a very small cost in order to substantially benefit many less fortunate children. According to Isserow, the fact that the parents' devotion would produce wrong action in different circumstances does not prevent us from judging their actual sacrifice as praiseworthy; this "strongly suggests to me", claims Isserow, "that judgments of moral praise do not stand or fall with judgments of counterfactual robustness" (2019: 263). Does the devoted parent case present a counterexample to the proposal I lay out here? I don't think so.

According to CRRT, for an agent to be praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons in the actual world, and for it to be the case that she would do the right thing for the relevant moral reasons in a range of normal counterfactual scenarios. To determine moral worth then, we ask whether *that same right action* would be performed, for the right reasons, in normal circumstances—the deontic status of the action is fixed across worlds. Whether and to what extent the devoted parents are praiseworthy depends upon whether they would perform the same sacrifice (conceived of as a right action) for the reasons which make it right in circumstances that are comparatively normal from an actual world perspective. Given that the parents are deeply devoted to their children, evidence suggests that they would continue to make the same right sacrifice in a range of normal worlds, in which case CRRT appeases Isserow's intuition that the parents are praiseworthy.

Notice that the counterfactual test employed by Isserow is different from the one employed by CRRT. Isserow considers whether the agent would continue to perform the action when circumstances make it morally wrong to do so. Whereas, I consider whether the agent would continue to perform the action in circumstances when it continues to be morally right to do so. Since the counterfactual tests are different, the two views produce different conclusions about moral worth. Isserow's complaint, therefore, does not target counterfactual accounts in general, but rather a specific counterfactual test. The objection has no grip on a view like CRRT which does not incorporate that test.

6 Conclusion

In this paper, I have tried to build on the success of RRT by supplementing it with a counterfactual framework. I have used the counterfactual apparatus as a way to measure the robustness of an agent's praiseworthy motive. In this way, counterfactuals have served as an epistemic tool, they have acted as a unifying, reliable and accurate proxy for denoting pertinent information about something which is constitutive of moral worth—motivational robustness. The truth value of the counterfactuals in and of themselves, I take it, is not something that endows an action with moral worth. In any case, I have argued that attending to an agent's counterfactual motives as well as their actual motives, allows us to successfully meet desiderata associated with theories of moral worth. Alongside this, I've argued that invoking counterfactuals means CRRT is able to better secure the non-accidentality constraint—a significant virtue if the problem of moral luck concerns you. And finally, by specifying that the counterfactuals relevant to moral worth are those instantiated in normal worlds, I take CRRT as able to appease the criticisms typically raised against modal accounts.

I wish to outline one final thought. I've argued that moral worth requires that (i) an agent does the right thing for the right reasons in the actual world and that (ii) the agent does the right thing for the right reasons in a range of normal worlds. Thus far I've taken (i) for granted in order to focus on defending (ii), but one might wonder whether taking (i) for granted is justified.[18] In particular, we might ask whether it is necessary for an agent to do the right thing from a response to the right reasons in the actual world to gain moral credit. Some cases would suggest not. Imagine that, as before, Lola is donating her £500 bonus to UNICEF from a praiseworthy motive. However, on this occasion, the actual world is not a normal world; moments before Lola clicks the mouse authorising the transaction, part of her ceiling falls in and kills her, she thus fails to satisfy condition (i). But imagine that Lola satisfies condition (ii). She possesses an incredibly robust praiseworthy motive which sees her donate in all normal worlds, including those where she needs to look for her lost debit card, or where she's anxious about a work project, or where she feels tired, grouchy or hungry. In fact, Lola's motive is so robust that she even manages to donate in less normal worlds, say worlds where she's bereaved after the sudden, tragic death of a family member. Despite the fact that Lola performs morally right counterfactual actions from an extraordinarily stable praiseworthy motive, CRRT says that Lola does not deserve moral credit because she fails to act well in *one* world—the actual world. Some will find this verdict counterintuitive. In particular, if Lola is denied praise, then the conditions for moral worth seem too demanding.

Perhaps this suggests that RRT isn't merely to be supplemented but to be overthrown. We might want to jettison (i) as a necessary condition for moral worth, making it the case that an agent is only required to act well in normal worlds from a response to the right reasons to gain moral credit. Although this will strike many as a radical position, notice that it is motivated in light of a conventional feature of moral worth—the non-accidentality constraint. Given that what happens in the actual world can sometimes depend upon accidents, we ought to doubt whether acting rightly in the actual world is always necessary for an action to have moral worth. More work needs to be done to explore the implications of such a view. But at the very least, cases such as this ought to prompt us to question a central assumption, pervasive in the moral worth literature, that an agent's behaviour in the actual world has a special claim to determining moral worth.

[18] I'm indebted to an anonymous reviewer for prompting me to think about this question and for their helpful discussion on the issue.

Acknowledgements I owe thanks to Robert Cowan and Neil McDonnell for their invaluable comments on earlier drafts of the paper. Thanks also to the two anonymous referees whose insights and thoughtful suggestions greatly improved the paper. Finally, I thank audiences at the University of Glasgow and the University of Oxford for their helpful discussions.
Funding No funds, grants, or other support was received.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Arpaly, N. (2002). *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press.

Arpaly, N., & Schroeder, T. (2013). *In praise of desire*. Oxford University Press.

Bennett, J. (1974). The conscience of huckleberry Finn. *Philosophy, 49*(188), 123–134.

Benson, P. (1987). Moral worth. *Philosophical Studies, 51*(3), 365–382.

Ferguson, B. (2012). Kant on duty in the groundwork. *Res Publica, 18*, 303–319.

Halpern, J. (2016). *Actual causality*. The MIT Press.

Henson, R. G. (1979). What Kant might have said: Moral Worth and the overdetermination of dutiful action. *The Philosophical Review, 88*(1), 39–54.

Herman, B. (1981). On the value of acting from the motive of duty. *The Philosophical Review, 90*(3), 359–382.

Isserow, J. (2019). Moral worth and doing the right thing by accident. *Australasian Journal of Philosophy, 97*(2), 251–264.

Kant, I. (1997). *Groundwork of the metaphysics of morals*. Cambridge University Press.

Lippert-Rasmussen, K. (2021). Praising without standing. *Ethics*. https://doi.org/10.1007/s10892-021-09374-2

Lord, E. (2017). What you're rationally required to do and what you ought to do (Are the Same Thing!). *Mind, 126*(504), 1109–1154.

Markovits, J. (2010). Acting for the right reasons. *Philosophical Review, 119*(2), 201–242.

Markovits, J. (2012). Saints, heroes, sages, and villains. *Philosophical Studies, 158*(2), 289–311.

Massoud, A. (2016). Moral worth and supererogation. *Ethics, 126*(3), 690–710.

McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies, 123*(1/2), 125–148.

Menzies, P. (2004). Difference-making in context. In N. Hall, L. Paul, & J. Collins (Eds.), *Causation and counterfactuals*. MIT Press.

Miller, D. J. (2018). Circumstantial ignorance and mitigated blameworthiness. *Philosophical Explorations, 22*(1), 33–43.

Montmarquet, J. (2012). Huck finn, aristotle, and anti-intellectualism in moral psychology. *Philosophy, 87*(1), 51–63.

Sliwa, P. (2016). Moral worth and moral knowledge. *Philosophy and Phenomenological Research, 93*(2), 393–418.

Smith, M. (2007). Ceteris paribus conditionals and comparative normalcy. *Journal of Philosophy of Logic, 36*, 97–121.

Smith, M. (2010). What else justification could Be1. *Noûs, 44*(1), 10–31.
Sorensen, K. (2010). Effort and moral worth. *Ethical Theory and Moral Practice, 13*(1), 89–109.
Stratton-Lake, P. (2000). *Kant, duty and moral worth.* Routledge.