Wrongness, Blameworthiness, and Overridingness

By Sam Mason

(Forthcoming in *Australasian Journal of Philosophy*)

According to the Overridingness Claim, if it is morally wrong for an agent to φ, then that agent has decisive normative reasons not to φ. A common argument for the Overridingness Claim appeals to the connection between moral wrongness and moral blameworthiness. I argue that this argument fails.

Keywords: Moral Wrongness, Moral Blameworthiness, Overridingness, Normative Reasons, Moral Normativity

1 Introduction

Consider the following two cases:

> *Fugitive Son*: A woman's son has committed a serious crime. She could hide him from the police, but if she does an innocent man will be wrongly convicted for the crime and sent to prison. (Wolf 2015: 41)

> *Williams's Gauguin:* Gauguin abandons his wife and children in Paris to go to Tahiti, where he can dedicate himself to painting. He produces brilliant work, work he would not have been able to produce had he stayed in Paris, and finds deep satisfaction and meaning in his artistic pursuits. The consequences for his wife and children, however, are dire: without their main source of financial support, they face great hardship and poverty. (Williams 1981: 22-26)

Both of these cases put pressure on *the Overridingness Claim*. This is the claim that if it is morally wrong for an agent to φ, then that agent has decisive normative reasons not to φ ('φ' stands for an action or omission).[1] At least on some ways of filling in further details, it seems clear that the mother acts morally wrongly if she hides her son from the police, and Gauguin acts morally wrongly in abandoning his wife and children in Paris. Yet for both cases there seems to be a further question as to whether these agents are justified from a comprehensive normative perspective. If the mother and Gauguin are justified from this perspective – if they have sufficient normative reasons to act in these ways, even though they act morally wrongly – then the Overridingness Claim is false.

[1] By 'moral wrongness', I mean all-things-considered moral wrongness, rather than *pro tanto* moral wrongness. This claim is sometimes called 'Moral Rationalism' (cf. Portmore 2011, 2021; Archer 2014). Since it is a contentious question what the relation between rationality and normative reasons is, I prefer to avoid this label. 'Moral Rationalism' would seem apt only if rationality consists in responding correctly to your normative reasons, but some philosophers deny this (e.g., Broome 2013).

The question of the relative importance of morality as compared with other normative domains, such as prudence, is intrinsically interesting. An additional reason for caring about the Overridingness Claim concerns the role this claim plays in some influential arguments against certain moral theories, including traditional (i.e., maximising, agent-neutral) act-consequentialism. Some philosophers argue against traditional act-consequentialism on the grounds that it conflicts with the Overridingness Claim (Stroud 1998: 171, 182-184; Portmore 2011: 29-32). According to this objection, agents do not always have decisive normative reasons to maximise agent-neutral value. Hence, given the Overridingness Claim, it cannot always be morally wrong for them to fail to do so. If the Overridingness Claim is false, objections of this kind will need to be rethought.

This paper considers and rejects what is perhaps the most important argument for the Overridingness Claim.[2] It remains neutral on whether the Overridingness Claim itself is correct or incorrect. The argument I reject starts from the claim that moral wrongness is (conceptually and/or metaphysically) analysable in terms of moral blameworthiness. A representative formulation of the analysis is as follows:

> *Moral Wrongness as Moral Blameworthiness (MB)*: It is morally wrong for an agent to φ iff (Def) φ-ing violates standards such that, if the agent violated those standards without a moral excuse, they would be morally blameworthy for violating them, where 'φ' stands for an action or omission.

[2] A presupposition of the Overridingness Claim is that good sense can be made of unsubscripted normative concepts. Subscripted normative concepts include MORALLY OUGHT and PRUDENTIAL REASON. These normative concepts concern what we ought to do or have a reason to do by the lights of a special normative domain. We might represent this with a subscript: OUGHT_{M} and REASON_{P}, where '$_{\text{M}}$' stands for morality and '$_{\text{P}}$' stands for prudence. It is a contentious question whether, in addition to subscripted normative concepts, there is good sense to be made of unsubscripted normative concepts, such as DECISIVE NORMATIVE REASONS. The Overridingness Claim presupposes, controversially, that good sense can be made of unsubscripted normative concepts, but I will not call this presupposition into question. Copp (1997) rejects this presupposition. See Dorsey (2016: Chapter 1) for a response.

Some philosophers have tried to derive the Overridingness Claim from MB and a further premise linking being morally blameworthy for φ-ing with having decisive normative reasons not to φ (Gibbard 1990: 299-300; Darwall 2006a: 97-99, 2006b: 292; Skorupski 2010: 295-301; Portmore 2011: 38-51, 2021: 51-62). I argue that this argument fails. My argument does not target MB but rather the premise linking being morally blameworthy for φ-ing with having decisive normative reasons not to φ.

Section 2 outlines the argument from MB to the Overridingness Claim. Section 3 argues that this argument is unsuccessful.

2 MB and the Overridingness Claim

I start by clarifying MB and explaining why it is an attractive view. For our purposes, the most important clarification concerns how moral blameworthiness is to be understood. Defenders of MB and related views have typically understood moral *blame* in terms of guilt, resentment, and indignation, and an agent's moral blame*worthiness* for φ-ing in terms of whether they would be a fitting target of guilt, resentment, and indignation for φ-ing (Gibbard 1990; Darwall 2006a; Kauppinen 2017). Both choices are well-motivated.

Blame is a diverse phenomenon: there are many kinds of reactions that can reasonably be considered forms of blame (cf. Shoemaker 2015; Rosen 2015: 67-68; Nussbaum 2016: 256-260). Among these reactions, guilt, resentment, and indignation have a good claim to being especially closely connected to moral wrongness. The main reason for this is their motivational profiles. Guilt characteristically motivates agents to make amends through such things as apologies and offers of compensation (Tangney and Dearing 2002: 19; De Hooge 2019). Complementarily, resentment and indignation characteristically motivate agents to make offenders hold themselves accountable by feeling guilty (for the right reasons) and making

amends for their conduct (Dill and Darwall 2014: 46-54).[3] Intuitively, there is a close connection between moral wrongness and making amends, and this makes it appealing for defenders of MB to understand moral blame in terms of these emotions.

Understanding moral blame*worthiness* in terms of whether agents would be *fitting* targets of moral blame is also well-motivated. A response is fitting when its object merits, or is worthy of, that response (Howard 2018). For example, admiration is fitting when it is felt towards something (proportionally) admirable, and fear is fitting when it is felt towards something (proportionally) dangerous. It is generally held that the fittingness of a response is independent of the value of the consequences of that response (D'Arms and Jacobson 2000; Rabinowicz and Rønnow-Rasmussen 2004). This is significant, since consequentialist accounts of moral blameworthiness would seem a poor match for MB, insofar as moral wrongness does not always line up with optimific moral blame.[4] At the same time, fittingness is not a specifically moral relation. Understanding moral blameworthiness in terms of the fittingness of moral blame, then, allows defenders of MB to avoid the extensional objections that would beset their view given a consequentialist account of moral blameworthiness, while deploying a normative relation that is sufficiently independent of moral wrongness to promise an illuminating analysis of it.

MB is an intuitively appealing view. It accounts for the close association between moral wrongness and making amends, given the motivational profiles of guilt, resentment, and indignation. Moreover, it can explain the common-sense idea that moral wrongs come in different degrees of seriousness. MB explains this in terms of differences in the degree of moral blameworthiness that attaches to different actions or omissions (cf. Gibbard 1990: 45). Finally,

[3] 'Motivate', as I use this term, does not imply that the agent in fact fulfils or even tries to fulfil the goal they are motivated to achieve. For example, an agent may be motivated to achieve a goal, but may not fulfil or even try to fulfil this goal if they are more strongly motivated to achieve a different, incompatible goal.

[4] However, see Miller (2014) for an argument that indirect consequentialist accounts of moral blameworthiness are compatible with MB.

another virtue of MB is that it can readily be generalised to account for the unity among different kinds of impermissibility (cf. McElwee 2017). On this approach, different kinds of impermissibility are to be explained in terms of different kinds of critical reactions that are fitting in response to them.

The argument from MB to the Overridingness Claim runs as follows. From MB it follows that:

> *The Wrongness-Blameworthiness Link*: If it is morally wrong for an agent to φ, then that agent would be morally blameworthy for φ-ing without a moral excuse.[5]

Now add the following claim connecting moral blameworthiness for φ-ing with having decisive normative reasons not to φ:

> *The Blameworthiness-Reasons Link*: If an agent would be morally blameworthy for φ-ing without a moral excuse, then that agent has decisive normative reasons not to φ.

An agent has *decisive* normative reasons not to φ if and only if not φ-ing is all-things-considered normatively required; they have *sufficient* normative reasons not to φ if and only if not φ-ing is all-things-considered normatively permissible. From the Wrongness-Blameworthiness Link and the Blameworthiness-Reasons Link, it follows that if it is morally wrong for an agent to φ, then that agent has decisive normative reasons not to φ (the Overridingness Claim).

[5] In principle, it is possible to accept the Wrongness-Blameworthiness Link without accepting MB. However, even if the Wrongness-Blameworthiness Link can be supported in some other way than by appealing to MB, my criticism of the argument from the Wrongness-Blameworthiness Link to the Overridingness Claim still goes through, since this criticism focuses on the Blameworthiness-Reasons Link.

3 The Blameworthiness-Reasons Link

The aim of this section is to undermine the role that the Blameworthiness-Reasons Link plays in the argument stated above. My aim is not to show that the Blameworthiness-Reasons Link is *false*. Rather, I aim to show that there is not a good argument *from* MB and the Blameworthiness-Reasons Link *to* the Overridingness Claim. As mentioned earlier, I remain neutral on whether the Overridingness Claim is correct. Now, if MB and the Overridingness Claim are true, then the Blameworthiness-Reasons Link is also true. But there is only a good argument from MB and the Blameworthiness-Reasons Link to the Overridingness Claim if the Blameworthiness-Reasons Link is plausible independently of the Overridingness Claim. Showing this would require giving an argument for the Blameworthiness-Reasons Link that does not depend for its plausibility on the Overridingness Claim. In this section, I consider and reject the two most promising arguments of this kind.[6] I argue that both of these arguments presuppose the Overridingness Claim at crucial points.[7]

3.1 Making Amends

In an early discussion of whether there is a good argument from MB to the Overridingness Claim, Allan Gibbard claims that there is an apparent incoherence involved in morally blaming

[6] An argument for the Blameworthiness-Reasons Link I will not consider is Darwall (2006b: 292). For a convincing response to this argument, see Dorsey (2020: 698-700).

[7] Dorsey (2016: 56-60) also argues that the argument from MB to the Overridingness Claim fails insofar as one of its premises relies for its plausibility on the Overridingness Claim, but he identifies the relevant premise as the Wrongness-Blameworthiness Link rather than the Blameworthiness-Reasons Link. He writes: 'without a prior commitment to [the Overridingness Claim], [the Wrongness-Blameworthiness Link] should simply be rejected' (2016: 56). (Dorsey seems to have changed his view in his (2020), insofar as he now argues that blameworthiness, including moral blameworthiness, need not be for actions we had decisive normative reasons not to perform). I argue below (in Section 3.1) that MB (which entails the Wrongness-Blameworthiness Link) is plausible independently of the Overridingness Claim, and can reasonably be accepted by someone who rejects the Overridingness Claim.

someone for φ-ing while being prepared to φ if in exactly like circumstances, and suggests, moreover, that this provides support for the Overridingness Claim:

> [Are] moral demands always demands of reason? Or can it sometimes make real sense to do things that are morally wrong?... …To judge that it fully makes sense to do a thing is, in effect, not to rule out doing it oneself, if in exactly like circumstances. Now, anger seems incoherent when joined to the thought "If I am in his shoes let me do the same". Likewise with guilt: it seems incoherent when joined to the thought "With the same opportunity, let me do it again". Anger and guilt seem indefensible at the very moment of embracing the act condemned. (1990: 299-300)

(Where Gibbard focusses on anger in general, I focus on resentment and indignation; where he focusses on whether it fully makes sense to do a thing, I focus on whether there are sufficient normative reasons for doing it.) Gibbard's argument has a couple of problematic features. First, it seems to move from the claim that moral blame would be *incoherent* under certain circumstances to the conclusion that it would not be *fitting* under those circumstances. But it is not clear how a claim about the incoherence of moral blame when it is combined with certain other attitudes could support the conclusion that moral blame is not merited by its object. Moreover, Gibbard does not justify or explain the claim that the apparent incoherence he points to is a genuine incoherence. Despite these problematic features, Gibbard's argument is worth investigating further. Doing so will allow us to develop a strong new argument for the Blameworthiness-Reasons Link that takes inspiration from Gibbard's argument while avoiding its problematic features.

We have seen that guilt characteristically motivates agents to make amends; complementarily, resentment and indignation characteristically motivate agents to make offenders hold themselves accountable by feeling guilty (for the right reasons) and making

amends for their conduct. At first glance, the following principle concerning making amends seems plausible:

> *The Repudiation Principle*: Making amends for φ-ing requires *repudiating* φ-ing – that is, committing to not φ-ing again if in exactly like circumstances (and communicating this commitment, in particular to your victim).

Being committed to not φ-ing again if in exactly like circumstances while being prepared to φ again if in exactly like circumstances seems incoherent. Insofar as guilt characteristically motivates agents to do something that requires them to undertake such a commitment, and resentment and indignation characteristically motivate agents to get someone to do something that requires them to undertake such a commitment, combining these emotions with being prepared to φ again if in exactly like circumstances would also seem to yield a tension or incoherence (albeit perhaps an incoherence of a weaker kind). So, if the Repudiation Principle is correct, then there seems to be a good case for claiming that the apparent incoherence that Gibbard points to in the passage quoted above is a genuine incoherence.[8]

But there is still a gap between claiming that it would be *incoherent* to feel an emotion in certain circumstances and claiming that it would not be *fitting* to feel it in those circumstances. Fortunately, we will not need to bridge this gap: there is a more direct way of arguing from the Repudiation Principle to the Blameworthiness-Reasons Link. To see this, we will need to look at the fittingness relation in more detail.

An influential account of fittingness holds that fittingness is a matter of *accurate representation*: for a response to be fitting is for it to involve an accurate representation of its object (cf. Graham 2014: 392-393; Rosen 2015: 70-71; Tappolet 2016: 87). Typically,

[8] Gibbard himself characterises guilt chiefly in terms of its connection with making amends (1990: 67-68, 139-140, 146-150).

defenders of this view take emotions to have normative representational content.[9] Plausibly, if emotions have such content, then among the factors that are relevant to determining what the content of a given emotion-type is are facts about the kinds of actions they typically motivate (cf. Portmore 2022: 53-54). For example, the claim that fear represents its objects as dangerous is plausible partly because fear typically motivates its subjects to become safe from its objects. Being motivated to become safe from X is an intelligible response to appraising X as dangerous. Now, if the Repudiation Principle is correct, it seems very likely that guilt, resentment, and indignation, insofar as they have normative representational content, include as part of this content a representation of the relevant agents as having had decisive normative reasons not to φ. Otherwise, it is hard to see how being motivated to make amends for φ-ing (or to get someone to feel guilty and make amends for φ-ing), where this involves committing to not φ-ing again if in exactly like circumstances, could be an intelligible response to the appraisals involved in these emotions. So, if fittingness is a matter of representational accuracy, and the Repudiation Principle is correct, there is a good case to be made for the claim that guilt, resentment, and indignation are fitting only if the relevant agent had decisive normative reasons not to φ. The Repudiation Principle therefore supports the Blameworthiness-Reasons Link, given an account of fittingness in terms of accurate representation.

Not everyone accepts that fittingness is a matter of representational accuracy.[10] Once this view is rejected, the main alternative views are that fittingness is analysable in terms of reasons (Cosker-Rowland 2019), or that it is a normatively primitive relation (McHugh and Way 2016; Howard 2019). On either of these views, there is a good case for claiming that the

[9] The literature on this issue is vast. For a few influential defences of the claim that emotions have normative representational content, see Nussbaum (2001), Roberts (2003), and Tappolet (2016). Gibbard (1990: 126-150) sketches two theories of emotions, on neither of which emotions essentially involve normative representational content.

[10] For criticisms, see, e.g., Svavarsdóttir (2014: 89-90, 101), Howard (2018: 6, 11), Naar (2021), and D'Arms (2022).

Repudiation Principle supports the Blameworthiness-Reasons Link. Whether a given emotion is a fitting response to its object, on either of these views, must depend partly on the nature of that emotion – including, surely, its motivational aspect. Now we can ask: what would the fitting objects of guilt have to be like, to merit a response that motivates agents to make amends, where this involves committing to not φ-ing again if in exactly like circumstances? And what would the fitting objects of resentment and indignation have to be like, to merit responses that motivate agents to get offenders to feel guilty and make amends? Plausibly, they would have to be φ-ings such that the agent had decisive normative reasons not to φ. If the agent had only *sufficient* reasons not to φ, or merely a *pro tanto* reason not to φ, then it is hard to see how they could be a fitting target of emotional responses that are tied to *repudiating* φ-ing.

So, it looks as though there is a good argument from the Repudiation Principle to the Blameworthiness-Reasons Link, whether fittingness is understood in terms of accurate representation, reasons, or else claimed to be normatively primitive:

Q1: Making amends for φ-ing requires repudiating φ-ing. (The Repudiation Principle)

Q2: If making amends for φ-ing requires repudiating φ-ing, then the Blameworthiness-Reasons Link holds.

C: The Blameworthiness-Reasons Link holds.

In the remainder of this subsection, I explore the Repudiation Principle further. My aim is not to argue that this principle is false. As with the Overridingness Claim and the Blameworthiness-Reasons Link, I remain neutral on whether the Repudiation Principle is correct. Rather, I aim to show that the plausibility of the Repudiation Principle depends on the plausibility of the Overridingness Claim. For there to be a good argument from MB and the Blameworthiness-Reasons Link to the Overridingness Claim, it would have to be possible to provide an argument for the Blameworthiness-Reasons Link that does not rely for its plausibility on the

Overridingness Claim. The argument for the Blameworthiness-Reasons Link from the Repudiation Principle does not meet this condition, because this principle depends for its plausibility on the Overridingness Claim.

What is involved in making amends? One element of making amends – arguably, the main element – is providing an apology. Besides apologising, another important element of making amends is offering compensation, where this is possible and appropriate. There does not seem to be any difficulty in understanding how someone could offer compensation if they do not repudiate φ-ing. For example, Gauguin in *Williams's Gauguin* could provide financial support to his family even if he stands by his decision. So, if the Repudiation Principle is correct, this must be because we cannot provide a full apology for φ-ing if we indicate that we are prepared to φ again if in exactly like circumstances.[11] Many find this an intuitively plausible condition on full apologies. For example, Nick Smith claims that 'categorical regret is essential to a full apology', where such regret 'entails a promise that the offender will not repeat the offense even under the same conditions and with the same incentives' (2005: 483). If this is right, then the Repudiation Principle must be correct.

But on further reflection, whether full apologies require that the apologiser repudiates φ-ing depends on whether the Overridingness Claim is correct. To see this, let us return to *Fugitive Son*.[12] Imagine that after a few years the son is discovered and the innocent man who

[11] What is a 'full' apology? Intuitively, apologies consist of several elements. Uncontroversially, these elements include admission of wrongdoing and acceptance of responsibility. An apology is full when all of the elements that make up apologies are present. It is worth emphasising that a full apology may be qualified in its content. For example, imagine that you tell someone a truth that they need to hear, but tell it in an unnecessarily harsh or tactless manner. Then you may offer a full apology for your tactlessness without apologising at all for telling the relevant truth.

[12] Let me clarify the roles that *Fugitive Son* and *William's Gauguin* play in my overall argument. It seems that, on some ways of filling in further details, the agents in these cases act morally wrongly, but it is not *obvious* that the agents lack sufficient normative reasons to act in these ways. Moreover, these cases are helpful in illustrating the claims I make in this section concerning the requirements of making amends. However, nothing in my overall argument depends on these cases in particular – any cases in which an action is morally wrong but favoured by normative reasons that are significant enough to call the Overridingness Claim into question could have served the same purposes.

was convicted in his place is released. The innocent man confronts the mother and demands an apology for hiding her son. She says something to the following effect: “I’m sorry for letting you be convicted for my son’s crime. I wronged you. If there was any way I could have prevented it from happening without giving him up, I’d have done it. But he asked me to shelter him and he’s my son after all”. The mother indicates that she stands by the action she performed. If in exactly like circumstances, she would do it again. Yet, consistently with this, she can express the judgments that it was a moral wrong for which she is morally responsible and that it mattered that she wronged him, and she can express her feelings of guilt for having performed it. Is this enough for her apology to be full? Intuitively, this depends on whether she had sufficient normative reasons for acting morally wrongly. If she did not, then she has not adequately acknowledged the importance of her moral wrongdoing. But if she did, then she has and her apology is full.

Some may not share this intuition, so it is worth providing further support for the claim that whether full apologies require repudiation depends on whether the Overridingness Claim is true. Further support is provided by reflection on the kinds of reparative work that apologies can do. One important kind is to withdraw threatening messages that are sent out by unexcused moral wrongdoing, such as the message that moral standards are unimportant and/or that victims of moral wrongdoing and their interests are unimportant (cf. Murphy 1982: 508-509; Hieronymi 2001: 548-549; Griswold 2007: 55-56). As Jeffrie Murphy writes: ‘[moral injuries] are ways a wrongdoer has of saying to us “I count and you do not,” “I can use you for my purposes,” or “I am here up high and you are there down below”’ (1982: 508). Apologies can withdraw these messages by indicating that their author does not stand by them. A further, related kind of reparative work that apologies can do is to repair damaged trust in others to respect moral standards going forward (Walker 2006: 191-229).

The mother's apology in *Fugitive Son* can do all these kinds of reparative work. It affirms the importance of moral standards and the importance of the victim and their interests (she indicates that, if there were any way she could have prevented the victim from being wrongfully convicted without giving up her son, she would have done this). It could also, in principle, help to repair damaged trust in her to respect moral standards going forward, by showing that she sets great importance by moral standards. Of course, the mother's apology does not demonstrate that she always gives *overriding* importance to respecting moral standards. But again, whether a full apology requires that one demonstrates this must surely depend on the relative importance of morality as compared with other normative domains. If the Overridingness Claim is false, and the mother has sufficient normative reasons to act morally wrongly, then her apology accurately registers the importance of moral standards and of her victim.

I conclude that whether the Repudiation Principle is plausible depends on whether the Overridingness Claim is plausible. Unless we have good antecedent reasons for accepting the Overridingness Claim, we do not have good reasons for accepting the Repudiation Principle. We therefore cannot appeal to the Repudiation Principle to show that the Blameworthiness-Reasons Link is plausible independently of the Overridingness Claim.

Before discussing a different argument for the Blameworthiness-Reasons Link, let me consider an objection to my criticism of the argument from the Repudiation Principle. The objection is that this argument need not be undermined by showing that the Repudiation Principle depends for its plausibility on the Overridingness Claim. What the argument could aim to show, it might be suggested, is that there is a cluster of connected claims – MB, the Blameworthiness-Reasons Link, the Repudiation Principle, and the Overridingness Claim – each of which is *prima facie* plausible on its own but which are mutually supportive in such a way that all of them gain additional plausibility. Understood in this way, it is not a problem for

the argument if the Repudiation Principle depends for its plausibility on the Overridingness Claim, since the point of the argument is to demonstrate the interdependence of this cluster of claims.

On this way of construing the argument, its ambitions are relatively modest. It aims to strengthen the confidence of those who already find the Overridingness Claim plausible, rather than derive this claim from a set of independently plausible premises. However, there are good reasons for thinking that the argument is unsuccessful in even this modest aim. The discussion of *Fugitive Son* suggests that MB, the Blameworthiness-Reasons Link, the Repudiation Principle, and the Overridingness Claim are not really an interdependent set of claims after all – for it is entirely possible to endorse MB while accepting different understandings of the connection between moral blameworthiness and normative reasons, the requirements of making amends, and the connection between moral wrongness and normative reasons. Moreover, we saw earlier that there is a raft of independent considerations favouring MB, including its intuitive appeal, its ability to explain the common-sense idea that moral wrongs come in different degrees of seriousness, and the fact that it can readily be generalised to account for the unity among different kinds of impermissibility. Hence, MB, the Blameworthiness-Reasons Link, the Repudiation Principle, and the Overridingness Claim are not really an interdependent set of claims, and we should reject the argument for the Blameworthiness-Reasons Link from the Repudiation Principle even on this re-interpretation of it.[13]

[13] See also D'Arms and Jacobson (2023: 205-207) for an account of the nature and fittingness conditions of guilt according to which guilt can be fitting even for actions that the agent had sufficient normative reasons for performing. This would cast some doubt on the Blameworthiness-Reasons Link. D'Arms and Jacobson do not accept MB, but do (as I read them) accept the claim that unexcused morally wrong actions merit guilt. This claim is closely related to the Wrongness-Blameworthiness Link. If D'Arms and Jacobson are right about the nature and fittingness conditions of guilt, then there are independent grounds to be skeptical about the Blameworthiness-Reasons Link, and to think that defending the Wrongness-Blameworthiness Link does not commit you to defending the Blameworthiness-Reasons Link.

3.2 Responsibility and Justice

In this subsection, I consider and reject a different argument for the Blameworthiness-Reasons Link. This argument is due to Douglas Portmore (2011: 47-51; 2021: 58-61). The argument hinges on the thought that an agent cannot be morally blameworthy for φ-ing when in φ-ing they flawlessly exercised a capacity in virtue of which they are a morally responsible agent in the first place. But if we reject the Blameworthiness-Reasons Link, we are forced to allow just this possibility. More carefully, Portmore's argument can be paraphrased, with some minor modifications, as follows:

> R1: An agent would be morally blameworthy for φ-ing without a moral excuse only if they have the relevant sort of control over φ-ing.
>
> R2: The agent has the relevant sort of control over φ-ing only if they have the capacity to respond appropriately to the relevant reasons.
>
> (From R1 and R2) R3: An agent would be morally blameworthy for φ-ing without a moral excuse only if they have the capacity to respond appropriately to the relevant reasons.
>
> R4: If an agent has sufficient normative reasons to φ, then, by flawlessly exercising their capacity to respond appropriately to the relevant reasons, they could be led to φ.
>
> R5: If an agent would be morally blameworthy for φ-ing without a moral excuse only if they have the capacity to respond appropriately to the relevant reasons, then they cannot be morally blameworthy for φ-ing without a moral excuse when, by flawlessly exercising this capacity, they could have been led to φ.

(From R3, R4, and R5) R6: An agent would not be morally blameworthy for φ-ing without a moral excuse if they have sufficient normative reasons to φ.

(From R6) R7: An agent would be morally blameworthy for φ-ing without a moral excuse only if they do not have sufficient normative reasons to φ.

R8: If an agent does not have sufficient normative reasons to φ, they have decisive normative reasons not to φ.

(From R7 and R8) C: If an agent would be morally blameworthy for φ-ing without a moral excuse, then that agent has decisive normative reasons not to φ. (The Blameworthiness-Reasons Link).

There is much in this argument that is worthy of discussion, but to keep the discussion manageable I will focus on just R5.[14]

What, positively, can be said in favour of accepting R5? Ultimately, Portmore seems to trace the plausibility of this premise, in my view rightly, to the thought that an agent is morally blameworthy for φ-ing only if it would not be unjust to morally blame them for φ-ing (2021: 58-61). He puts the point as follows:

> …it seems that appropriate blame cannot be unjust, and yet it would be unjust to hold an agent responsible on the condition that they have the capacity to be guided by sound practical reasoning and then blame them for acting as they might very well be led to act

[14] Other criticisms of Portmore's argument for the Blameworthiness-Reasons Link can be found in Dorsey (2020: 700-701), Tucker (2022: 2037-2040), and Ventham (2023: 13-14). All of these criticisms focus on how the phrase 'the relevant reasons' is to be understood. While these philosophers press their criticisms in different ways, the core complaint seems to be that, if R4 is to be plausible, 'the relevant reasons' must be interpreted as meaning 'all of the normative reasons bearing on the decision whether to φ'. But these philosophers argue that, if 'the relevant reasons' is understood this way, R2 is implausible or question-begging. My criticism of Portmore's argument takes a different tack, focussing on the role played by considerations concerning the *justice* of moral blame. I aim to show that philosophers who reject the Overridingness Claim can nonetheless accept that justice is a constraint on moral blameworthiness, in the sense that an agent is morally blameworthy for φ-ing only if it would not be unjust to morally blame them for φ-ing.

> if they are guided by sound practical reasoning. And since an agent can be led to perform any act that they have sufficient reason to perform when guided by sound practical reasoning, it seems inappropriate to blame them for freely, attributively, and knowledgeably doing what they have sufficient reason to do, as this would be unjust. (2021: 58-59)

In other words, it seems unjust for agents to be morally blamed for acting in ways they could have been led to act by flawlessly exercising a capacity in virtue of which they are eligible for moral blameworthiness in the first place. If an agent is morally blameworthy for φ-ing only if it would not be unjust to morally blame them for φ-ing, it follows that agents cannot be morally blameworthy for φ-ing when φ-ing could have resulted from flawlessly exercising such a capacity. For example, suppose that the only way that Tom can spare himself a minor embarrassment is by telling a small lie to Mary. And suppose further that he has sufficient normative reasons either to lie or refrain from lying. Appreciating that he has sufficient normative reasons to do either of these things, Tom chooses to lie. Would it be just for Mary and others to morally blame Tom for lying? Intuitively, this would be unjust. Tom has exercised his capacity to respond appropriately to the relevant reasons flawlessly, so morally blaming him seems unjust (cf. Portmore 2021: 60-61).[15]

In the remainder of this subsection, I argue that concerns about the justice of moral blame do not in fact support R5. Given this, R5 is unsupported and so we do not have good reasons to accept the argument for the Blameworthiness-Reasons Link outlined above. To show that concerns about the justice of moral blame do not support R5, there are three routes we

[15] The claim that it would be unjust to morally blame agents for performing actions they have sufficient normative reasons to perform has some intuitive force even if we reject Portmore's claim that moral blameworthiness requires the capacity for sound practical reasoning. Guilt involves suffering, and resentment and indignation aim, in part, at inducing guilt in offenders. How could these emotions be justly directed towards agents who act in ways they have sufficient normative reasons to act? To be sure, Portmore's claim that moral blameworthiness requires the capacity for sound practical reasoning strengthens the force of the concern that morally blaming such agents would be unjust, but the concern also has some intuitive force independently of Portmore's claim.

could take. First, it might be argued that justice simply does not apply to moral blame: like other emotional responses (such as sadness and fear, perhaps), moral blame might simply not be apt for assessment as just or unjust (cf. Smith 2019). Second, it might be argued that justice is not a constraint on moral blameworthiness: it is not the case that an agent is morally blameworthy for φ-ing only if it would not be unjust to morally blame them for φ-ing (cf. Vargas 2004: 225). For example, imagine that an agent culpably performs a morally wrong action, but later suffers a very great misfortune. Perhaps it could be unjust to morally blame them even though they are morally blameworthy. Third and finally, we might try to meet the concerns about the justice of morally blaming the relevant agents on their own terms. That is, we might concede that moral blame *is* apt for assessment as just or unjust, and that justice *is* a constraint on moral blameworthiness, but argue that it need not be unjust to morally blame people for actions they could have been led to perform by flawlessly exercising their capacity to respond appropriately to the relevant reasons. I will pursue the third approach, but it is worth highlighting that, even if this approach proves unsuccessful, R5 may still be unsupported due to one of the other two approaches sketched above.

Note first that, if R5 is to be part of a non-question begging argument for the Overridingness Claim, it must be unjust to morally blame someone for φ-ing even under the following circumstances: they could have been led to φ by flawless practical reasoning, but φ-ing was nonetheless morally wrong and they met the conditions of moral responsibility in φ-ing. If it would not be unjust to morally blame someone for φ-ing under these circumstances, then R5 depends for its plausibility on the Overridingness Claim (since, if the Overridingness Claim is true, then the circumstances as described are impossible). Now, there is a good argument from MB and the Blameworthiness-Reasons Link to the Overridingness Claim only if it is possible to provide an argument for the Blameworthiness-Reasons Link that does not rely for its plausibility on the Overridingness Claim. If R5 depends for its plausibility on the

Overridingness Claim, then Portmore's argument fails to meet this condition. I will now argue that under the hypothetical circumstances described above it would not be unjust to morally blame someone for φ-ing. Hence, the argument for the Blameworthiness-Reasons Link outlined above is unsuccessful.

So far, in considering the justice of morally blaming people for actions they could have been led to perform by flawless practical reasoning, I have focussed exclusively on the perspective of potential blamees. Focusing on this perspective reveals a complaint which, if undefeated, seems to sustain a charge of injustice. The complaint is:

> *Potential Blamees' Complaint*: In performing an action they could have performed as a result of flawless practical reasoning, the potential blamee could have flawlessly exercised one of the capacities in virtue of which they are eligible for moral blameworthiness in the first place. But then it seems unjust for them to be morally blamed for φ-ing.

This complaint provides some support for the charge that it is unjust to morally blame people for actions they could have been led to perform by flawless practical reasoning. However, this complaint is defeated under the hypothetical circumstances described in the previous paragraph: cases in which the relevant agent could have been led to φ by flawless practical reasoning, but φ-ing was nonetheless morally wrong and they met the conditions of moral responsibility in φ-ing.

Note, first, that we might think that the fact that the agent is a culpable moral wrongdoer is already enough to defeat the charge that it would be unjust to blame them morally. But there is also a further point to be made. As we began to see in the last subsection, unexcused moral wrongdoing characteristically imposes various harms on victims and the wider community that are most effectively alleviated insofar as culpable moral wrongdoers are held morally

accountable. Unexcused moral wrongs typically send out threatening messages, such as the message that moral standards are unimportant and/or that victims of moral wrongdoing and their interests are unimportant. Relatedly, unexcused moral wrongs typically undermine trust in others to respect moral standards going forward. Finally, unexcused moral wrongs also typically harm victims in other ways as well. Apologies and further reparative gestures such as compensation, and the demands for these forms of reparation that are the natural expressions of resentment and indignation, are the most effective means of repairing many of these harms that we have at our disposal. Apologies and compensation withdraw threatening messages by indicating that their author does not stand by them, and help to restore damaged trust by indicating that the apologetic agent sets great importance by moral standards (even if not always overriding importance). Moreover, these reparative effects are greatly strengthened insofar as demands for apologies and compensation are made by victims and the wider community.

These points about moral blame and the harms typically imposed by unexcused moral wrongdoing help to provide a general explanation of why, at least for the most part, morally blaming unexcused moral wrongdoers is just. The key point is that it is just for the costs of repairing the harms of unexcused moral wrongdoing to fall on the wrongdoer, since they brought about these harms without a moral excuse and they are best placed to repair them through apologies and further reparative gestures such as compensation. Moral blaming emotions are justly directed towards unexcused moral wrongdoers insofar as these emotions aim at ensuring that these agents repair the harms of unexcused moral wrongdoing through holding themselves accountable, feeling guilty, and making amends.

This general explanation of why, at least for the most part, it is just to morally blame unexcused moral wrongdoers applies in the hypothetical circumstances under consideration: cases in which the relevant agent could have been led to φ by flawless practical reasoning, but

φ-ing was nonetheless morally wrong and they met the conditions of moral responsibility in φ-ing. As unexcused moral wrongdoings, these actions have various characteristic harms, including sending out threatening messages, undermining trust, and, typically, harming victims in other ways as well. As we saw in the previous subsection, even agents who acknowledge that they had sufficient normative reasons for performing an action can be committed to repairing many of the harms it caused, by offering sincere apologies and compensation. It seems just for the costs of repairing the harms of unexcused moral wrongdoing to fall on the wrongdoer even in the hypothetical circumstances at issue, and so morally blaming such agents seems just. It would surely be unjust for victims and members of the wider community to bear by themselves the costs of repairing the harms imposed by the agent's unexcused moral wrongdoing. Even if culpable moral wrongdoers have some complaint against being morally blamed in such cases, then, this complaint is defeated.

I conclude that R5 in the argument for the Blameworthiness-Reasons Link stated above is not plausible independently of the Overridingness Claim, and hence that we do not have good reasons for accepting this argument.

4 Conclusion

An influential argument for the Overridingness Claim appeals to the Wrongness-Blameworthiness Link (if it is morally wrong for an agent to φ, then that agent would be morally blameworthy for φ-ing without a moral excuse) and the Blameworthiness-Reasons Link (if an agent would be morally blameworthy for φ-ing without a moral excuse, then that agent has decisive normative reasons not to φ). I have argued that this argument assumes what it sets out to establish. The plausibility of the Blameworthiness-Reasons Link depends on the plausibility of the Overridingness Claim. Unless we have good antecedent reasons for

accepting the Overridingness Claim, we do not have good reasons for accepting the Blameworthiness-Reasons Link.[16]

References

Archer, A. (2014). Moral Rationalism without Overridingness. *Ratio, 27*(1), 100-114.

Broome, J. (2013). *Rationality Through Reasoning.* Oxford: Wiley-Blackwell.

Copp, D. (1997). The Ring of Gyges: Overridingness and the Unity of Reason. *Social Philosophy and Policy, 14*(1), 86-106.

Cosker-Rowland, R. (2019). *The Normative and the Evaluative: The Buck-Passing Account of Value.* Oxford: Oxford University Press.

D'Arms, J. (2022). Fitting Emotions. In C. Howard, & R. Cosker-Rowland (Eds.), *Fittingness* (pp. 105-129). Oxford: Oxford University Press.

D'Arms, J., & Jacobson, D. (2000). Sentiment and Value. *Ethics, 110*(4), 722-748.

D'Arms, J., & Jacobson, D. (2023). *Rational Sentimentalism*. Oxford: Oxford University Press.

Darwall, S. (2006a). *The Second-Person Standpoint: Morality, Respect, and Accountability.* Cambridge: Harvard University Press.

Darwall, S. (2006b). Morality and Practical Reason: A Kantian Approach. In D. Copp (Ed.), *The Oxford Handbook of Ethical Theory* (pp. 282-320). Oxford: Oxford University Press.

De Hooge, I. E. (2019). Improving Our Understanding of Guilt by Focusing on Its (Inter)personal Consequences. In B. Cokelet, & C. Maley (Eds.), *The Moral Psychology of Guilt* (pp. 131-148). London: Rowman & Littlefield.

Dill, B., & Darwall, S. (2014). Moral Psychology as Accountability. In J. D'Arms, & D. Jacobson (Eds.), *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics* (pp. 40-83). Oxford: Oxford University Press.

Dorsey, D. (2016). *The Limits of Moral Authority*. Oxford: Oxford University Press.

Dorsey, D. (2020). Respecting the Game: Blame and Practice Failure. *Philosophy and Phenomenological Research, 101*(3), 683-703.

Gibbard, A. (1990). *Wise Choices, Apt Feelings.* Cambridge: Harvard University Press.

Graham, P. A. (2014). A Sketch of a Theory of Moral Blameworthiness. *Philosophy and Phenomenological Research, 88*(2), 388-409.

Griswold, C. (2007). *Forgiveness.* Cambridge: Cambridge University Press.

[16] I'm very grateful to Brad Hooker, Gerald Lang, Andrew Mason, Douglas Portmore, Joe Saunders, Pekka Väyrynen, and several anonymous referees for helpful feedback on earlier drafts, and to an audience at the University of Leeds for useful discussion.

Hieronymi, P. (2001). Articulating an Uncompromising Forgiveness. *Philosophy and Phenomenological Research, 62*(3), 529-555.

Howard, C. (2018). Fittingness. *Philosophy Compass, 13*(11).

Howard, C. (2019). The Fundamentality of Fit. *Oxford Studies in Metaethics, 14*, 216-236.

Kauppinen, A. (2017). Sentimentalism, Blameworthiness, and Wrongdoing. In K. Stueber, & R. Debes (Eds.), *Ethical Sentimentalism* (pp. 107-132). Cambridge: Cambridge University Press.

McElwee, B. (2017). Supererogation Across Normative Domains. *Australasian Journal of Philosophy, 95*(3), 505-516.

McHugh, C., & Way, J. (2016). Fittingness First. *Ethics, 126*(3), 575-606.

Miller, D. E. (2014). "Freedom and Resentment" and Consequentialism: Why 'Strawson's Point' Is Not Strawson's Point. *Journal of Ethics and Social Philosophy, 8*(2), 1-23.

Murphy, J. G. (1982). Forgiveness and Resentment. *Midwest Studies in Philosophy, 7*(1), 503-516.

Naar, H. (2021). The fittingness of emotions. *Synthese, 199*(5-6), 13601-13619.

Nussbaum, M. (2001). *Upheavals of Thought.* Cambridge: Cambridge University Press.

Nussbaum, M. (2016). *Anger and Forgiveness: Resentment, Generosity, Justice.* Oxford: Oxford University Press.

Portmore, D. (2011). *Commonsense Consequentialism.* Oxford: Oxford University Press.

Portmore, D. (2021). *Morality and Practical Reasons.* Cambridge: Cambridge University Press.

Portmore, D. (2022). A Comprehensive Account of Blame: Self-Blame, Non-moral Blame, and Blame . In A. Carlsson (Ed.), *Self-Blame and Moral Responsibility* (pp. 48-76). Cambridge: Cambridge University Press.

Rabinowicz, W., & Rønnow-Rasmussen, T. (2004). The Strike of the Demon: On Fitting Pro-attitudes. *Ethics, 114*(3), 391-423.

Roberts, R. (2003). *Emotions.* Cambridge: Cambridge University Press.

Rosen, G. (2015). The Alethic Conception of Moral Responsibility. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The Nature of Moral Responsibility: New Essays* (pp. 65-88). Oxford: Oxford University Press.

Shoemaker, D. (2015). *Responsibility from the Margins.* Oxford: Oxford University Press.

Skorupski, J. (2010). *The Domain of Reasons.* Oxford: Oxford University Press.

Smith, A. M. (2019). Who's Afraid of a Little Resentment? *Oxford Studies in Agency and Responsibility, 6*, 85-111.

Smith, N. (2005). The Categorical Apology. *Journal of Social Philosophy, 36*(4), 473-496.

Stroud, S. (1998). Moral Overridingness and Moral Theory. *Pacific Philosophical Quarterly, 79*(2), 170-189.

Svavarsdóttir, S. (2014). Having Value and Being Worth Valuing. *Journal of Philosophy, 111*(2), 84-109.

Tangney, J. P., & Dearing, R. L. (2002). *Shame and Guilt.* New York: Guilford Press.

Tappolet, C. (2016). *Emotions, Values, and Agency.* Oxford: Oxford University Press.

Tucker, C. (2022). Too far beyond the call of duty: moral rationalism and weighing reasons. *Philosophical Studies, 179*, 2029-2052.

Vargas, M. (2004). Responsibility and the Aims of Theory: Strawson and Revisionism. *Pacific Philosophical Quarterly, 85*(2), 218-241.

Ventham, E. (2023). The Division of Normativity and a Defence of Demanding Moral Theories, *Ethical Theory and Moral Practice, 26*, 3-17.

Walker, M. U. (2006). *Moral Repair.* Cambridge: Cambridge University Press.

Williams, B. (1981). *Moral Luck.* Cambridge: Cambridge University Press.

Wolf, S. (2015). *The Variety of Values.* Oxford: Oxford University Press.