

Free Will and the Asymmetrical Justifiability of Holding Morally Responsible

Benjamin Vilhauer, William Paterson University of New Jersey

Abstract

This paper is about an asymmetry in the justification of praising and blaming behavior which free will theorists must acknowledge even if they do not follow Wolf and Nelkin in holding that praise and blame have different control conditions. That is, even if praise and blame have the same control condition, one must have stronger reasons for believing that it is satisfied to treat someone as blameworthy than to treat someone as praiseworthy. Blaming behavior which involves serious harm can only be justified if the claim that the target of blame acted freely cannot be reasonably doubted. But harmless praise can be justified so long as the claim that the candidate for praise did not act freely can be reasonably doubted. Anyone who thinks the debate about whether we have free will is truth-conducive has to acknowledge that reasonable doubt is possible in both these cases.

Introduction

One of Susan Wolf's contributions to the free will literature is to highlight the intuition that we must meet a higher standard of justification to blame than to praise.¹ Wolf points out this justificatory asymmetry in the course of arguing for what might be called an ontological asymmetry in the control conditions of praise and blame. (On her view, people can be blameworthy only if they could have acted differently, but can be praiseworthy even if they could not have acted differently.) There has been relatively little discussion of the justificatory asymmetry. Conversations have tended to focus on the claim that there is an

¹ "Asymmetrical Freedom" (*Journal of Philosophy* 77.3, 1980). Wolf does not speak in exactly these terms, but I take this to be one of her points.

ontological asymmetry (a claim which has been further defended by Dana Nelkin in recent papers²) and the justificatory asymmetry has largely been treated as a side-topic in those conversations.³ But the claim that there is a justificatory asymmetry is largely independent of the claim that there is an ontological asymmetry, and it has important implications for free will theory even if we do not accept the claim that there is an ontological asymmetry. (In this paper, "free will" means whatever satisfies the control condition of moral responsibility.⁴) For example, the claim that there is a justificatory asymmetry implies that even if we accept a theory on which the control condition is the same in the contexts of praise and blame, we must have better reasons to believe that the theory is true to appeal to it to justify treating someone as blameworthy than to appeal to it to justify treating someone as praiseworthy. Even if we accept a Frankfurt-style compatibilism⁵ about the control condition in the contexts

² Nelkin, "Responsibility, Rational Abilities, and Two Kinds of Fairness Arguments" (*Philosophical Explorations* 12.2, 2009, 151-165) and "Responsibility and Rational Abilities: Defending an Asymmetrical View" (*Pacific Philosophical Quarterly* 89, 2008, 497-515).

³ One exception is Gary Watson's "Two Faces of Responsibility" (*Philosophical Topics* 24, 1996, 227–248) but he considers different issues than I do here.

⁴ These terms are used in a broad sense meant to cover all the various accounts of the control condition of moral responsibility, including strong libertarian notions such as dual control, and weaker compatibilist notions such as guidance control or hierarchical control. For guidance control, see John Martin Fischer and Mark Ravizza, *Responsibility and Control: a Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998). For hierarchical control, see Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68.1 (1971).

⁵ Frankfurt holds that alternative possibilities are not necessary for moral responsibility. See

of both praise and blame, this intuition seems to imply that we must have better reasons to believe that Frankfurt-style compatibilism is true to appeal to it to justify treating someone as blameworthy than to justify treating someone as praiseworthy. The same would appear to hold if we instead accept a dual-control libertarian account of the control condition in the contexts of both praise and blame. My goal in this paper is to develop this point, and then to draw some tentative conclusions about when we can justify holding people morally responsible. The idea is to see how far this project can be carried without making independent commitments on the other issues that typically occupy the attention of free will theorists (such as whether compatibilism is true and whether free will requires alternative possibilities or agent causation).

This paper has six sections. In the first, I discuss the intuition about the asymmetric justifiability of praise and blame in more detail. In the second, I explain why the asymmetric justifiability of praise and blame implies an asymmetry in the justificatory standards that must be met by claims about free will in the contexts of praise and blame. I then ask: how much higher is the standard for blame? In sections three and four I offer a partial answer. Seriously harmful blame can only be justified if the claim *that the target of blame acted freely* cannot be reasonably doubted. On the other hand, harmless praise can be justified as long as the claim *that the candidate for praise did not act freely* can be reasonably doubted. (I regret the tortured language in which these two claims are expressed, but precision demands it.) In section 5, I argue that anyone who thinks the debate about whether we have free will is truth-conducive must acknowledge that both these claims can be reasonably doubted. Finally, in the sixth, I consider a couple of potential objections and reply to them.

“Alternate Possibilities and Moral Responsibility”, *Journal of Philosophy* 66 (1969), 829-839.

The debate about whether we have free will is of course not the only important debate about free will. For example, compatibilists and libertarians, who agree that human beings have free will, have a debate about whether free will is compatible with determinism. There may well be compatibilists and libertarians who think that debate is truth-conducive but do not think the debate about whether we have free will is truth-conducive. I will make no attempt to argue that people who do not think that the debate about whether we have free will is truth-conducive must accept that the reasonable doubt standard is met. But many people do seem to think that the debate about whether we have free will is truth-conducive, so I think it is worthwhile to set out this argument.⁶

1. The Asymmetry Intuition

This paper is about an asymmetry in the standards for blaming behavior (actions used to express the Strawsonian negative reactive attitudes) and praising behavior (actions used to express the positive reactive attitudes). To be concise, I will often use "blame" to refer to the former, and "praise" to refer to the latter. In other words, as used here, "blame" and "praise" involve not just *believing* that someone is morally responsible, but acting in such a way as to *hold* him morally responsible. (I do not mean to claim that blaming and praising are essentially matters of behaving in certain ways—this is merely an expository aid.)

⁶ Explaining what it is for a doubt to be reasonable is a notoriously difficult matter. For this reason, I will not try to set out a context-independent definition of reasonable doubt and then argue that everyone should have doubts about the claims at issue that satisfy this definition. I will only argue that people who think the debate about whether we have free will is truth-conducive must acknowledge that there is reasonable doubt about these claims.

The intuition that praise and blame are asymmetrically justifiable can be explained as part of a more general intuition that harms and benefits are asymmetrically justifiable. Justice demands that arguments for harming people be to be held to a higher standard than arguments for refraining from harming them or benefiting them. All philosophers must acknowledge that this asymmetry exists, though disagreement is to be expected when it comes to giving a detailed explanation of why it exists. All will probably agree that, in one way or another, the primary purpose of morality has to do with getting people to benefit and refrain from harming each other. If this is right, then arguments for harming people run against the primary purpose of morality (in at least a *prima facie* way), while arguments for benefiting people run with it. In my own view, this asymmetry exists because of personhood-based desert claims which can be made by or on behalf of all people. In other words, it exists because of facts about how people deserve to be treated just because they are people. People deserve to be given the benefit of the doubt. They deserve not to be intentionally harmed unless there is a very strong justification for that harm. They also deserve to be benefited when someone can do so without significant harm to anyone. When someone is considering providing such a benefit, the burden of justification is much lighter: that is, if all parties to an interaction agree that something is a benefit which can be provided without significant harm to anyone, no further justification is typically expected.

Praise and blame are commonly kinds of intentional benefit and harm. That is, praise and blame commonly produce valuable or harmful experiences for the people who are their recipients, and this effect is commonly intended by the people doing the praising or blaming. This is the case even when praise and blame do not involve the obvious benefit and harm of tangible reward and punishment. That is, praise obviously involves intentional benefits when it includes a tangible reward, as when a rich rescuee lavishes not just words of gratitude but a monetary reward upon the brave rescuer. But when a poor rescuee only offers words of

gratitude, he may still intentionally benefit the rescuer, by intentionally causing the rescuer to have a valuable emotional experience of his own heroism, nobility, self-worth, etc.

Something similar holds for blame: when blame includes retributive punishment by (e.g.) imprisonment or death, it obviously includes intentional harm, but even if it just involves the expression of a condemning attitude, this may intentionally cause the target of blame to experience emotional pain. In light of this, the asymmetry in the justificatory standards for harm and benefit would appear to imply an asymmetry in the justificatory standards for harmful blame and beneficial praise.

It may be that not all praise is or is intended to be beneficial, and that not all blame is or is intended to be harmful. There may be cases where we praise or blame people solely as a way of getting them to reflect on what they have done, for example, without intending or causing them to have valuable or harmful experiences. I am not sure that we can sensibly call such behavior praise and blame, but for present purposes let us suppose that we can. The arguments of this paper would imply no asymmetry in the justificatory standards for such behavior. But it seems clear that a lot of praise is or is intended to be beneficial, and that a lot of blame is or is intended to be harmful. So the asymmetry is relevant for a good bit of the praise and blame that actually goes on.

Beneficial praise and harmful blame are special kinds of benefit and harm, since they can only be legitimate if the people who are their recipients deserve them based on how they have acted. That is, if people ever really deserve praise or blame, then they deserve it not just because they are people, but because of how they have acted, and people can only deserve things based on their actions if they are morally responsible for their actions.

My own view is that we need to distinguish between personhood-based desert and action-based desert to explain this difference. (The idea is that all acts of harm and benefit must conform to personhood-based desert, but praise and blame are special in that they must

also conform to another kind of desert, action-based desert.⁷) For purposes of this paper, however, the details of how one explains the difference are probably less important than seeing the difference that needs explaining.

⁷ Some philosophers think that all desert is action-based. Examples include Rachels ("What People Deserve", *Justice and Economic Distribution*, ed. J. Arthur and W.H. Shaw, Englewood Cliffs, NJ: Prentice-Hall, 1978, p. 157) and Sadurski (*Giving Desert Its Due: Social Justice and Legal Theory*. Dordrecht: D. Reidel, 1985, p. 131). Smilansky holds a related position, i.e., that giving up the belief that human beings are morally responsible for their actions implies giving up all our morally significant beliefs about desert ("Responsibility and Desert: Defending the Connection." *Mind* 105.417 (1996), pp. 157-63). It is not absurd to suppose that actions are the only kind of desert base, since the category of action-based desert claims is very broad. Yet there are desert claims that can only be supposed to be action-based with great difficulty. It is natural to think that people deserve respect, access to their rights, equal treatment before the law, not to be used as a mere means to others' ends, and to be given the benefit of the doubt, and that they deserve these things not because of facts about her actions, but simply because they are people. I argue that personhood-based desert has a role to play in free will theory in [author's paper 1]. Some may think it is a mistake to call this "desert", and may prefer instead to call it "entitlement". I think that the terminology of desert is better than the terminology of entitlement for the personhood-based claims at issue here. "Entitlement" is sometimes used for claims that are in a deep sense morally arbitrary but still legitimately enforceable in a shallow sense. (For example, some ethicists might say that a wealthy farmer could be entitled to all the food grown on his lands even if his field hands were malnourished.) But the personhood-based claims at issue in this paper are in no sense morally arbitrary.

Some may wonder whether the claims about moral responsibility involved in justifications of praise and blame make praise and blame so different from other kinds of harm and benefit that the justificatory asymmetry relevant for other kinds of harm and benefit has no bearing on justifications for praise and blame. But I cannot see why this should be so. It is a straightforward matter to include claims about moral responsibility within the scope of the justificatory asymmetry. We can hold justifications for blame to a higher standard than justifications for praise by holding all the claims that play roles in justifications for blame to a higher standard than the claims that play roles in justifications for praise. Since claims about moral responsibility play roles in justifications for both praise and blame, those claims must be held to a higher standard when they appear in justifications for blame than when they appear in justifications for praise.

How much higher is the justificatory standard for claims about moral responsibility in justifications for blame than in justifications for praise? If the praise/blame asymmetry exists because of the benefit/harm asymmetry, then it seems natural to suppose that the standards for both praise and blame are slopes that rise as the harm they cause increases. But it is beyond my reach in this paper to survey this whole range. My argument here will focus on the standards for two kinds of limiting cases at the extremes of this range: seriously harmful blame, and praise that can be given without significantly harming anyone.

It is probably clear enough how blame can be seriously harmful. Blame can include retributive bodily violence. Blame can also include the retributive infliction of great emotional pain as punishment for bad actions, or the sort of shaming behavior that permanently marks its target with an indelible stigma of monstrosity or absurdity, and such

emotional violence can sometimes be even more harmful than bodily violence.⁸ Other kinds of seriously harmful blame include retributive punishment by execution, and retributively justified imprisonment under dreadful conditions that corrode personality and make dignified living very difficult (like those that prevail in contemporary prisons).⁹ To keep things concise, I will call seriously harmful blame "serious blame" from now on. (Probably only a small minority of instances of blame amount to serious blame, and my argument here only applies to these instances. This restricts the argument's applicability, but given the ethical importance of these instances, I do not think this does much to diminish its significance.)

In talking about praise which can be given without harming anyone significantly, I have in mind praise that benefits the recipient and doesn't cause any significant harm as a by-product. (I will call such praise "harmless praise".) Harmless praise does not make the person giving the praise miss a significant opportunity to do something else worthwhile. It also does not harm third parties in any significant way. For example, if we select one person from a group of people to praise her for doing something good, this can cause the other people in the group to feel sorrowful about not having done well enough to merit recognition. The valuable experience of the person singled out comes at the expense of the painful experiences of the others. But there are instances where harms like these could be avoided.

⁸ In my view, not all remorse is based on self-retribution. In [author's paper 2], I argue that there is a kind of remorse which is based on suffering in sympathy with the person one has harmed rather than on self-retribution.

⁹ Punishment does not have to be retributively justified, and the argument of this paper is only relevant for retributive justifications of punishment. For a recent alternative, see Pereboom's quarantine justification in his book *Living Without Free Will* (Cambridge: Cambridge University Press, 2001).

We can sometimes single out individuals privately so that others' feelings are not hurt. On other occasions, we can be egalitarian with praise. Suppose everyone in the group has tried to do good things. We might praise them all for trying.¹⁰ It may well be that everyone who can be hurt by being excluded from praise sometimes tries to do good things. (That is, it may well be that everyone sometimes tries to do good things except sociopaths who do not care whether or not they are praised.) If this is right, then we might maximize the amount of harmless praise we give by praising everyone we meet who sometimes tries to do good things.

2. Asymmetry and Free Will

To be morally responsible, we must satisfy the conditions of moral responsibility. There is broad agreement that there are multiple conditions of moral responsibility: a control condition, as well as conditions having to do with agents' understanding of their situation, and their motivation for acting as they do. Only the control condition is at issue in this paper.

¹⁰ Egalitarian praise may not be harmless when people who achieve more than others protest that they deserve to be singled out for praise in a way that excludes the others. I will not take a position on the legitimacy of this protest here. I am suspicious of the claim that egalitarian praise is ever unjust, and I am inclined to see sorrow caused by not being included in praise as more ethically important than sorrow caused by having to share praise with others. But there may be a real problem about cheapening the practice of praising in some cases, and instances of praise which would cheapen the practice of praising would not be harmless. For present purposes, I only wish to claim that, in cases where people who achieve more than others choose not to protest egalitarian praise (perhaps out of magnanimity), they have not been treated unjustly, and there is no obstacle to harmless praise. (Also see note 10 below.)

But for my purposes here, I must draw further distinctions between three kinds of control-related conditions:

- (I) the control condition of moral responsibility;
- (II) the condition of a justified *belief* that someone satisfies the control condition;
- (III) the condition of being justified in *treating someone* as satisfying the control condition of moral responsibility.

(I) is satisfied by anyone who acted freely. (II) is satisfied not by the fact that someone acted freely, but rather by a justified belief that he acted freely. (III) is also satisfied by a justified belief, but I will argue that it is different in the contexts of serious blame and harmless praise. Asymmetry in (I) seems to imply asymmetry in (II), and asymmetry in (II) seems to imply asymmetry in (III). But asymmetry in (III) does not imply asymmetry in (I).

As mentioned at the outset, Wolf and Nelkin argue for asymmetry in (I). On their view, alternative possibilities are necessary to satisfy the control condition of blameworthiness, but not praiseworthiness. Such an asymmetry in (I) would imply an asymmetry in (II). That is, if alternative possibilities are necessary to satisfy the control condition for blameworthiness but not praiseworthiness, then a justified belief that someone had alternative possibilities is necessary for a justified belief that he satisfies the control condition of blameworthiness but not praiseworthiness. This in turn would imply an asymmetry in (III). If a justified belief that someone had alternative possibilities is required for a justified belief that he satisfies the control condition of blameworthiness but not praiseworthiness, then we presumably need some reason to believe that he had alternative possibilities to treat him as satisfying the control condition of blameworthiness, but not to treat him as satisfying the control condition of praiseworthiness.

Wolf's and Nelkin's view that there is an asymmetry in (I) has been discussed with great interest in the literature. But in this paper, I will not take a position on this view. I will instead argue that there can be asymmetry in (III) without asymmetry in (I).

Suppose, just for the sake of argument, that there is no asymmetry in (I), that is, that the very same kind of control satisfies the control condition in the contexts of praise and blame. Can we have asymmetry in (II) in these circumstances? Our answer to this question depends upon how we understand beliefs about moral responsibility. We may think that our beliefs are open to moral critique even when we are not actually acting upon them, and that we therefore ought to hold our beliefs about moral responsibility to a higher standard in the context of blameworthiness than in the context of praiseworthiness. But if "ought" implies "can", then this seems to imply that we can exercise voluntary control over what we believe. Some oppose such doxastic voluntarism.

Philosophers who accept views like those of P.F. Strawson and Jay Wallace can probably accept doxastic voluntarism about beliefs in moral responsibility without raising any new problems for their views.¹¹ To speak roughly, such philosophers think that believing that someone satisfies the control condition implies readiness to apply the reactive attitudes and practices to him, so long as the other conditions of moral responsibility are satisfied. In other words, to believe that someone satisfies the control condition is to have decided to treat him as satisfying the control condition. Such philosophers presumably want to be voluntarists about deciding how to act, so they should probably also be voluntarists about

¹¹ See P.F. Strawson, "Freedom and Resentment" (*Proceedings of the British Academy* 48, 1963, pp. 1-25) and also R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press, 1996). Section 6 discusses a potential objection to my argument based on another aspect of the Strawson/Wallace approach.

beliefs about the control condition. Doxastic voluntarists can hold that there is asymmetry in (II) even without asymmetry in (I), since their view gives them room to hold that we should voluntarily hold our beliefs about agents' satisfaction of the control condition to a higher standard in the context of blameworthiness than in the context of praiseworthiness. Since asymmetry in (II) implies asymmetry in (III), they can also accept the possibility of asymmetry in (III) without asymmetry in (I).

Opponents of doxastic voluntarism (doxastic involuntarists) should also accept the possibility of asymmetry in (III) without asymmetry in (I), though for different reasons. Suppose they hold that, irrespective of the moral context at hand, to justifiably believe that something is the case, we must have reasons to think that there is a greater than 50% chance that it is the case. (This is controversial, of course, but this argument does not turn on the precise number chosen: it could easily range between 50% and 75%, for example.) Suppose once again that there is no asymmetry in (I), and that the kind of control which satisfies the control condition in both contexts is control of kind C. Doxastic involuntarists should hold that (II) is symmetrical, and that under the assumptions just stated, reason to think there is a greater than 50% probability that someone possessed C justifies the belief that he possessed C in the contexts of both praise and blame. But if (III) is satisfied in the context of serious blame only if there is reason to think that the probability that the target possessed C is a lot higher than 50%, then it can fail to be satisfied in cases where (II) is satisfied. Suppose (as will be argued below in section 3) that (III) is satisfied for serious blame only if the claim *that the target acted freely* cannot be reasonably doubted. In a case where the target of blame had an 80% probability of possessing C, doxastic involuntarists should hold that (II) is satisfied, but not (III). Now consider harmless praise. Suppose (as will be argued below in section 4) that (III) is satisfied for harmless praise as long as the claim *that the candidate did not act freely* can be reasonably doubted. In a case where the candidate for harmless praise had a

30% probability of possessing C, doxastic involuntarists should hold that (III) is satisfied, but not (II).¹²

3. Serious Blame and Reasonable Doubt

In this section, it will be argued that (III) is satisfied for serious blame only if the claim *that the target acted freely* cannot be reasonably doubted.¹³ Call this the "serious blame principle". If (III) is not satisfied for serious blame, then serious blame is not justified, and we are obligated to refrain from it. So the serious blame principle implies that we are obligated to refrain from serious blame whenever it can be reasonably doubted that the target of the blame had free will with respect to the action at issue.

The serious blame principle is supported by an intuition which can be drawn out by considering the following scenario. Suppose that bad neuroscientists coercively implant a device into Skip's brain which can randomly cause him to do terrible things. It does not just give extra "oomph" to intentions to act which Skip forms based on his native practical reasoning abilities (that is, the abilities which developed in him in the normal, species-typical way). It provides what I will call a "complete system" of capacities for practical reasoning: identification of reasons for action (bad reasons, in this case), formation of desires, intentions, and volitions, and whatever else one might think necessary in a complete process of practical

¹² For another helpful discussion of the distinction between true desert and being justified in treating people as deserving, see Norvin Richards' "Luck and Desert" (*Mind* XCV.378, 1986, pp. 198-209.)

¹³ The analogy to the criminal conviction standard made in this section is drawn from a longer argument about retribution and reasonable doubt which I present in [author's paper 3]. I recapitulate it here in order to connect it to the more general point about asymmetrical justifiability which is the focus of the present paper.

reasoning that terminates in action. The complete system provided by the implanted device is qualitatively different from Skip's native complete system—that is, it does not merely duplicate the sort of reasons-identification, intention-formation, etc. that Skip's native complete system is disposed to provide. If it operates, Skip's native system is totally bypassed—that is, the practical reasoning capacities provided by the device are the only ones which play a role in the explanation of the action. There is no way for anyone to know if it will operate, and if it does, there is no way for anyone to know that it has done so. (The practical reasoning that the device can cause would be seamlessly integrated into Skip's conscious experience in such a way that even he would be unable to tell that it had not originated in him in the ordinary way.) After the device is implanted, Skip commits a grisly murder. Good neuroscientists discover the device and promptly remove it, and everyone in the scenario can be sure that the device operated either only once, or not at all. All free will theorists should agree that if the device operates, then Skip does not act freely—that is, that he does not satisfy (I).

Would it be justified for Skip to receive serious blame for the murder? (In the terms used earlier, does Skip satisfy (III) in the context of serious blame?) For example, would we be justified in retributively inflicting intense suffering on him? In the scenario as described, presumably not. What if we could know that there was a 90% probability that the device did not operate? I think we would still hold that serious blame was not justified. What if we could know that there was a 95% or 99% probability that the device did not operate? It would be sensible to be very careful around Skip, but I think we would still hold that serious blame would not be justified. This line of argument raises the bar until we arrive at the reasonable doubt standard. Only if the claim *that the device did not operate* could not be reasonably doubted could serious blame be justified.

Some may object that the intuitions drawn out by this scenario are not clear enough to offer unequivocal support to the serious blame principle. I am not sure that this is right, but the argument for the serious blame principle does not have to rest on this scenario alone. It gets some extra support from an analogy with another "reasonable doubt" principle which is widely recognized to be a requirement of justice, that is, the principle observed in criminal legal proceedings that the accused can only be convicted of a crime if it is proven beyond reasonable doubt that he acted criminally. The conviction standard and the serious blame principle are both grounded on the same basic intuition about justice. The intuition is really just a further specification of the intuition described earlier about the asymmetrical justifiability of harm and benefit. Justice requires arguments for harming people to be held to a higher standard than arguments which are not for harming anyone, and it requires arguments for seriously harming people to be held to an especially high standard: there must be no room for reasonable doubt about their soundness. This holds whether the harm at issue is blame or of some other kind.

In courts of law, an argument that someone has committed a crime is typically part of a larger argument that that person is to be given a punishment which will cause serious harm such as imprisonment under morally and emotionally corrosive conditions, or even death, in some countries. This is why justice demands that we hold arguments that someone has committed a criminal act to the "reasonable doubt" standard. By contrast, in civil trials, the most common sort of penalty involves payment of monetary damages, which typically does not involve serious harm, and in that context, the burden of proof is the lower "preponderance of the evidence" standard, which is typically understood to require a demonstration that there is a greater than 50% probability that the defendant acted as accused. (It is probably the case that some civil penalties are more harmful than some criminal penalties—paying monetary damages can be more harmful than a few months in jail if they are high enough. But this can

be chalked up as one of the many ways in which the current legal system fails to embody our intuitions about justice.)

When a claim about free will serves as a premise in a justification for serious blame, it makes sense to hold it to the same standard as arguments in the criminal courtroom. That is, in this context, the claim that someone has free will plays a role in an argument for serious harm, just as the claim that someone has committed a crime typically does. This suggests that the serious blame principle has the same justification as the criminal conviction standard.

Start here

4. Harmless Praise and Reasonable Doubt

In this section, it will be argued that (III) is satisfied for harmless praise so long as the claim *that the candidate for praise did not act freely* can be reasonably doubted. Call this the "harmless praise principle". The harmless praise principle does not imply an obligation in the way that the serious blame principle does. The harmless praise principle merely says that (III) is satisfied for harmless praise if it can be reasonably doubted that the candidate did not act freely. If (III) is satisfied for harmless praise, along with whatever other conditions there may be, then harmless praise is justified. But presumably the fact that acting in some way is justified does not imply that acting in that way is obligatory.

The harmless praise principle is supported by an intuition which can be brought out by considering the following scenario. Suppose that confusedly public-spirited neuroscientists coercively install a device into Pip's brain which can randomly cause him to do heroic things. It is just like Skip's device, except the reasons-identification capacities it provides identify the right sort of reasons for action, rather than bad reasons. It provides a complete system, and if it operates, Pip's native complete system is totally bypassed. There is no way for anyone to know if it will operate, and if it does, there is no way for anyone to know that it has done so. The practical reasoning which the device can cause would be

seamlessly integrated into Pip's conscious experience in such a way that even he would be unable to tell that it had not originated in him in the ordinary way. Pip goes on to rescue Doug from a hungry shark despite an awareness that he is risking his own life, motivated by a belief that Doug needs help and a desire to provide that help. Different neuroscientists promptly discover the device and remove it despite some consequentialist misgivings, and everyone in the scenario can be sure that the device operated no more than once.

All free will theorists should agree that if the device operated, then Pip did not satisfy (I), that is, the control condition for moral responsibility. But this may require more argument in the Pip case than it did in the Skip case, in light of the view held by Wolf and Nelkin that (I) is asymmetrical, and that the requirements agents must meet to satisfy (I) are weaker in the context of praise than in the context of blame. In my view, all defensible theories of free will must endorse what we might call the "minimal mechanism ownership requirement". This requirement says that agents cannot satisfy (I) in any context unless their own capacities for practical reasoning play some role in the explanation of the action at issue. If an agent is manipulated through the coercive implantation of a complete system which (a) is qualitatively different from that agent's native complete system, (b) operates no more than once, and (c) can randomly and totally bypass the agent's native system without his consent, then the agent does not own the implanted system. As explained earlier, the total bypass means that the practical reasoning capacities provided by the device are the only ones which play a role in the explanation of the action. This is of course what happens in the Pip case if the device operates. Since the only practical reasoning capacities which play a role in the explanation of the action do not belong to Pip, Pip's own capacities play no role in the explanation, so Pip does not meet the minimal mechanism ownership requirement.

Since we cannot know whether the device operated, there is good reason to doubt that Pip satisfied the control condition of praise. But it nonetheless seems justifiable to give Pip

harmless praise. Suppose that we could know that there was a 10% chance that the device did not operate. Would we be inclined to give him harmless praise then? I think so. How about a 5% or 1% chance? Pip deserves the benefit of the doubt, and since harmless praise is the only thing at stake, it is not clear that we could have a reason to accept 10% as good enough, but not 5% or 1%. Even though the grounds for thinking that Pip acted freely eventually become quite weak as we proceed with this line of thought, they seem to remain strong enough, given the context. This line of argument lowers the bar until we arrive at the reasonable doubt standard. No reasonable person would suppose that Pip could be justifiably praised if it could not be reasonably doubted that the device operated. But since the praise is harmless, it seems justifiable to praise him as long as it can be reasonably doubted that the device operated. Put differently, it seems justifiable to praise him as long as the claim *that he did not act freely* can be reasonably doubted. If this is right, then the harmless praise principle is correct.¹⁴

The argument for the serious blame principle got some extra support from an analogy with the criminal conviction standard. But there is nothing in jurisprudence which might play

¹⁴ Suppose that someone else, who has no such device, was praised for rescuing Doug from a hungry shark last week. He might be pained if Pip got the same kind of praise he got—he might object that there is more reason to doubt that Pip acted freely than there is to doubt that he acted freely, so it is unfair to give Pip the same kind of praise he got. I have my doubts that this would be unfair. But as far as the argument of this paper is concerned, I am willing to concede that praise would not be harmless in this case. It is enough to point out that last week's rescuer might be magnanimous enough not to be pained or to object, and that in this case, it would be possible to give our Pip the same kind of praise last week's rescuer got without being unfair. (Also see note 8 above.)

a similar role in the argument for the harmless praise principle. Whatever extra support we can give this argument must be based on more general philosophical considerations.

As a first step, it may be worth pointing out that if reasonable doubt standards govern both serious blame and harmless praise, then there is some symmetry within the asymmetry. (III) would have different branches in the contexts of serious blame and harmless praise, but the branches would be symmetrical, at least in a sense. It makes sense to try the simplest theory first, so it makes sense to try to preserve some symmetry within the asymmetry.

This may seem to come at the price of proposing a peculiarly low standard for harmless praise. But remember that harmless praise is a kind of harmless benefit. When all parties agree that an action under consideration is a harmless benefit, we usually expect no further justification for it.

Harmless praise cannot be supposed to be entirely typical in this regard, of course. Harmless benefits as such are justified merely by the claim that they are harmless benefits, so people who have no reason to doubt this claim have no reason to ask for any further justification. But harmless praise must be seen to rest on a further claim, that is, a claim that the candidate for praise is morally responsible for the action at issue. If one doubts this further claim, it makes sense to ask for a defense of it, and part of the defense must involve giving some reason to suppose that the candidate acted freely.

But when we ask how high a justificatory standard to apply to the claim that the candidate acted freely, it makes sense to look at the moral context in which the question is asked. It seems important that the context is an attempt to provide someone with a harmless benefit. If the primary purpose of morality has something to do with getting people to benefit and refrain from harming one another, then the context is an attempt to do something that helps fulfill the primary purpose of morality. I think this makes it clear that the justificatory standard should be low.

But even if it is clear that the standard should be low, it may not be clear precisely how low it should be. The reasonable doubt standard built into the harmless praise principle is the lowest possible standard. (Anything lower would amount to no standard at all.) Why should we accept the lowest possible standard instead of a standard which is higher, but still low? Consider the following point. In the context of harmless praise, it does no harm to use the lowest possible standard, and it may do some harm to use a higher standard. That is, using a higher standard may deprive some candidates for harmless praise of benefits they might otherwise have.

Some will no doubt reply that if the correct standard is higher than reasonable doubt, then there would indeed be some harm in accepting the harmless praise principle, that is, the harm of accepting a false belief. But the most obvious way in which a false belief can be harmful is by prompting people to act in harmful ways, and mistakenly accepting the harmless praise principle could not have this effect. It may be that we can also be harmed by accepting a false belief even if it has no impact on our actions. The sheer fact of not seeing things as they are may be harmful. But surely the harm of this kind which could arise from mistakenly accepting the harmless praise principle is quite slight. Some philosophical errors might involve significant harm of this kind, for example, believing that God exists if this is false. But the point at issue is small and peripheral by comparison. So it seems fair to assume that the harm of depriving some candidates for praise of benefits they might otherwise have is greater than the harm of mistakenly accepting the harmless praise principle. (I must emphasize that I am not arguing that we should not care about getting things right when it comes to the harmless praise principle, only that if we are not sure precisely how low the justificatory standard should be, then we should err on the side of reducing harm.)

There are a couple of points which it may be helpful to reiterate at this juncture. First, in claiming that harmless praise of Pip is justified, I am not claiming that anyone is obligated

to give him harmless praise. As far as my argument here is concerned, we may have the right to adopt a policy of praising only those who have undeniably satisfied the conditions of moral responsibility in doing truly great deeds. We may even have the right to model ourselves on the proud man of the *Nicomachean Ethics* 4.3, who seems to offer praise rarely if at all. All that I am arguing here is that when we properly understand harmless praise as a kind of harmless benefit, we will see that anyone who is inclined to praise Pip can legitimately do so, as long as it can be reasonably doubted that the device operated.

Second, it must be borne in mind that it is only the control condition of praise that is at issue. I am not arguing that it is justifiable to praise Pip if we doubt that he knew the rescue was dangerous. (I have stipulated that Pip knows it is dangerous.) Neither am I arguing that it is justifiable to praise Pip if we doubt whether he had a laudable motivation for the rescue or was instead motivated by a desire for a reward. (I have stipulated that Pip is only motivated by a desire to help.)

In the next section, I will argue that no matter what theory of free will we hold, if we think that a debate about whether someone acted freely is truth-conducive, we must accept that it can be reasonably doubted both that he did act freely, and also that he did not. If this is right, and if the serious blame and harmless praise principles are correct, then it follows that anyone who thinks a debate about whether someone acted freely is truth-conducive must accept that he does not satisfy condition (III) for serious blame, and that he does satisfy it for harmless praise.

5. Reasonable Doubt About Free Will

In this section, it will be argued that no matter what theory of free will one holds, if one thinks that a debate about whether someone acted freely is truth-conducive, one must

accept that it can be reasonably doubted both that he did act freely, and also that he did not.¹⁵

It may be helpful to begin with a few terms for describing debates. Some debates can be represented as focused on a central claim. A debate about whether someone acted freely is an example. Its central claim is that this person (i.e. a person of such-and-such a description) acted freely in this situation (i.e. a situation of such-and-such a description). Debates which are focused on a central claim include a pro side and a con side. The pro and the con side each have a basic position. The basic position of the pro side is that the central claim is true. The basic position of the con side is that the central claim is false.

As these terms are to be understood here, one can be on the pro or con side of a debate without believing it to be truth-conducive. One person might hold that the central claim is true, and another might hold that it is false, but they might agree that the debate about it is a waste of time. The first of these people would be on the pro side, and the second would be on the con side, but neither would see the debate as truth-conducive. Someone of a different temperament might take a side and might value the debate, but as bracing intellectual exercise and nothing more. To believe a debate to be truth-conducive is to believe that working on at least some of the objections to one's basic position posed by the opposite side either tends to prompt one to revise the theory one uses to support one's basic position in a way that makes it more likely to be true, or helps one to better understand why one's existing theory is true.

Now that the terms have been explained, the argument can proceed quickly. Objections can only be truth-conducive if they are reasonable, and reasonable objections are grounds for reasonable doubts about the basic positions to which they are objections. This means that, whether one is on the pro or con side of a debate, if one takes the debate to be

¹⁵ This argument is a more general version of an argument I describe in [author's paper 3].

truth-conducive, one must accept that it is possible to reasonably doubt one's basic position. So, in the case of a debate about whether someone acted freely, those on the pro side must accept that it is possible to reasonably doubt the claim that he acted freely, and those on the con side must accept that it is possible to reasonably doubt the claim that he did not act freely.

Two points of clarification may be helpful before the implications of this argument for free will are discussed in more detail. First, I am not claiming that we can only get closer to a true theory, or gain a better understanding of why our existing theory is true, when we are working on reasonable objections. One can of course be struck by a good idea at any time, even when one is working on unreasonable objections. The claim intended here is the weaker claim that, since unreasonable objections do not direct our attention to features of our theory which are reasonably thought of as implausible, there is nothing about unreasonable objections *as such* which would justify us in believing that working on them would be truth-conducive. We might suppose that they could haphazardly cause us to get closer to a true theory, or to a better understanding of why an existing theory is true, but not that they would *lead* us to these outcomes.

Second, I am not claiming that everyone who thinks a debate is truth-conducive must actually be able to doubt his own basic position in the debate. It is common for people to become so deeply committed to their basic positions that it becomes psychologically impossible for them to doubt them. But this is no objection to this argument, because the fact that it is psychologically impossible for some people to doubt a claim does not imply that it cannot reasonably be doubted. I do not even claim that *if* the debaters were reasonable, *then* they would doubt their basic positions. Philosophical disagreement is a complicated matter. As far as this argument is concerned, there may be what Richard Feldman calls "mutually recognized reasonable disagreement" between the pro and con sides. That is, it may be that

even when both sides recognize that the other side can reasonably doubt their basic position, they can remain reasonable without doubting their own basic position.¹⁶ To defend this argument, it is enough to claim that those on both sides would not be unreasonable if they came to doubt their own basic positions, and it seems fair to claim this much.

Philosophers will disagree about when we can have truth-conducive debates about whether someone acted freely. The cases most congenial to free will are ones with normal adults acting in normal conditions—call such cases "normal cases". I can imagine extreme Strawsonians who might hold that the belief that people often act freely in normal cases is so fundamental to the meaning of discussions about free will that those discussions lose all sense if we put it in question (though I am not sure anyone has actually taken this position). Such Strawsonians might therefore hold that debating about whether people in normal cases ever act freely is not truth-conducive. But this would presumably be a minority view. Others may be committed to varieties of reductive physicalism which make the idea of free will look too absurd for a debate about it to be truth-conducive even in normal cases. But this too would be a minority view.

Anyone who thinks it is truth-conducive to debate whether people in normal cases act freely must accept that it can be reasonably doubted that they do. If it is true that normal cases are the situations where free will is most likely to be found, then it seems to follow that if it can be reasonably doubted that people act freely in normal cases, then it can be reasonably doubted that anyone ever acts freely. If the serious blame principle is correct, this implies that anyone who believes a debate about normal cases to be truth-conducive must

¹⁶ See Richard Feldman, "Epistemological Puzzles about Disagreement", in S. Hetherington (ed.) *Epistemology Futures* (New York: Oxford University Press, 2006).

accept that (III) is never satisfied for serious blame, and therefore that serious blame is never justified.

Anyone who thinks it is truth-conducive to debate whether people act freely in normal cases must also accept that it can be reasonably doubted that they do *not* act freely. If the harmless praise principle is correct, this implies that anyone who thinks this debate is truth-conducive must accept that people in normal cases satisfy (III) in the context of harmless praise.

Non-normal cases do not provide ideal conditions for finding free will, but even non-ideal conditions may sometimes provide room for a truth-conducive debate. There are lots of cases where there may not be room for a truth-conducive debate about free will, for example, cases where we know the agent did something by accident or because of an irresistible compulsion, and cases involving very young children or people with very profound cognitive or behavioural disabilities. But there is probably room for truth-conducive debate in cases involving people who do things because of seemingly resistible compulsions, and cases involving older children or people with less profound cognitive or behavioural disabilities. If we think debates are truth-conducive in these cases, then we must accept that (III) is satisfied for harmless praise in these cases as well.

6. Objections and Replies

In this section I will consider two potential objections to the arguments presented above. The first is as follows. The argument of the previous section implies that anytime we think it is truth-conducive to debate a proposition, we must acknowledge that it can be reasonably doubted, and some might object to this. For example, it implies that anyone who thinks that the debate about other minds skepticism is truth-conducive must accept that it is possible to reasonably doubt that other minds exist. But it is clearly absurd to claim that just

because someone thinks this debate is truth-conducive, she should act as if the existence of other minds has been put in doubt. In my view, it is true that anyone who thinks the debate about other minds skepticism is truth-conducive must acknowledge that the existence of other minds can be reasonably doubted. But as mentioned earlier, it seems possible to accept that some claim can be reasonably doubted without doubting it ourselves. And the possibility of reasonable doubt about the existence of other minds certainly would not imply that we should act as if other minds do not exist. It would only imply this if we accepted a principle which said that the possibility of reasonable doubt about the existence of other minds requires us to act as if other minds do not exist, and I cannot see any reason to accept such a principle. This highlights an important difference between doubts about free will and doubts about other minds. If the argument of this paper is right, morality requires us to treat reasonable doubt about free will as practically significant. But it is hard to see how morality could require us to treat reasonable doubt about the existence of other minds as practically significant. If the primary purpose of morality has something to do with getting people to benefit one another, or at least getting them to refrain from harming one another, then morality seems to imply a commitment to the existence of other minds.¹⁷

The second objection I want to discuss is based on a concern about whether (I) and (III) are really as independent as I have argued. The sort of independence I have claimed to find may seem to presuppose a metaphysical view of moral responsibility according to which the properties agents must possess to satisfy the control condition must be characterizable independent of our practices of holding agents morally responsible. To presuppose this would be to reject, without argument, the influential views of P.F. Strawson and Jay Wallace, who (roughly speaking) define moral responsibility as appropriate accessibility to the reactive

¹⁷ I make the same point in [author's paper 3].

attitudes and the practices which express them.¹⁸ (Let us call views like this "deflationary theories".) But I do not take myself to be presupposing this. In my view, any defensible deflationary theory must allow for cases where agents are in a deep sense appropriately accessible to reactive attitudes and practices, but nobody is justified in treating them as appropriately accessible, because nobody has strong enough evidence. It must also allow for cases where agents are in a deep sense *not* appropriately accessible, but everyone is justified in treating them as appropriately accessible, because everybody has strong enough evidence. Not to allow for such cases would be to slide into what I take to be an unacceptable subjectivism about moral responsibility. Consider a modified version of the Skip scenario. Suppose that Skip has all the abilities that deflationary theorists require for appropriate accessibility to the reactive attitudes and practices, and that he commits a murder without his implanted device operating. Suppose everybody in the scenario knows that that there was a 50% probability of its operating, but suppose that as a matter of fact it does not operate, though nobody in the scenario knows this. I think deflationary theorists ought to say that Skip satisfies the control condition for being appropriately accessible to the reactive attitudes and practices, but that nobody in the scenario is justified in treating him as appropriately accessible. In other words, deflationary theorists ought to acknowledge that in this scenario, (I) is satisfied, but (III) is not. We can modify the Pip scenario in a similar way. Suppose that everybody in the scenario knows that Pip's device has a 50% chance of operating, and that as a matter of fact it *does* operate, though nobody knows this. In this case, I think deflationary theorists ought to hold that Pip does not satisfy the control condition for being appropriately accessible, but that everybody in the scenario is justified in treating him as appropriately accessible. In other words, (I) is not satisfied, but (III) is.

¹⁸ These philosophers are also mentioned earlier, in section 2.

Conclusion

To conclude, I would like to emphasize a point which is implicit in the above discussion, but which may be worth making explicit: it would seem that free will believers and free will skeptics can accept the argument presented here without giving up their basic positions. By "free will skeptics", I mean those who hold that the control condition of moral responsibility is never satisfied, and by "free will believers", I mean those who hold that it is sometimes satisfied. It seems consistent to hold both that some claim is true and also that it can be reasonably doubted. If this is right, then free will skeptics could accept that it can be reasonably doubted that candidates for harmless praise do not act freely. If they accept the harmless praise principle, they should conclude that harmless praise is justifiable (at least in cases where there is no reason to doubt that any conditions for moral responsibility which do not involve control are satisfied). Similarly, free will believers could accept that it can be reasonably doubted that targets of serious blame act freely, and if they accept the serious blame principle, then they should conclude that serious blame is never justified. This would allow skeptics and believers to preserve their basic positions while meeting each other halfway on some important issues.