Hypocritical Blame as Dishonest Signalling

Adam Piovarchy

The University of Notre Dame, Australia

This paper proposes a new theory of the nature of hypocritical blame and why it is objectionable, arguing that hypocritical blame is a form of dishonest signaling. Blaming provides very important benefits: through its ability to signal our commitments to norms and unwillingness to tolerate norm violations, it greatly contributes to valuable norm-following. Hypocritical blamers, however, are insufficiently committed to the norms or values they blame others for violating. As allowing their blame to pass unchecked threatens the signaling system, our strong interest in maintaining valuable norm-following by tracking who has what commitments justifies objecting to hypocritical blame. This theory has a number of strengths over competing accounts: it delivers intuitive verdicts about when blame is objectionable across a range of cases, it is a naturalistic explanation, it is consistent with a leading theory of the nature of blame, it explains why hypocritical pronouncements that don't feature blame are similiarly objectionable, it does not rely on contentious analyses of the nature of 'standing', and it preserves the common intuition that hypocrites are in some way dishonest.

Keywords: blame, hypocrisy, hypocritical blame, standing, signalling.

1. Introduction

What is objectionable about hypocritical blame? A full answer to this question requires two things:

1) An account of *when* hypocritical blame is objectionable that accords with our blaming practices.
2) A convincing explanation of *why* hypocritical blame is objectionable.

We currently lack any account which convincingly meets both of these criteria. This paper argues that a *Commitment Account of Hypocritical Blame*, once developed, can do so. By combining it with a *Costly Signal Theory of Blame*, and showing how our interest in maintaining blame's power to credibly signal our commitments can generate prescriptions for our blaming practices, this paper provides a comprehensive account of why hypocritical blame is objectionable.

2. Commitments and Blame

Hypocritical blame, and the question of what makes it objectionable, have received a lot of recent attention (Bell 2013; Fritz & Miller 2018, 2019; 2022; Herstein 2017, 2020; Isserow & Klein 2017, Lippert-Rasmussen 2021; Piovarchy 2020a, 2023a; Riedener 2019; Roadevin 2018; Rossi 2018, 2020; Tierney 2021; Todd 2019). Objectionable blame per se is nothing new; it is inappropriate to blame people who did not do anything wrong, for instance. But since hypocritical blame often targets culpable wrongdoers, and being a culpable wrongdoer is traditionally thought to make one blame*worthy,* the puzzle is accounting for why facts about blam*ers* could be relevant to the 'legitimacy' of their blame. As many have noted, we have very particular objections to hypocritical blame, such as 'Who are *you* to blame *me*?' or 'Look who's talking!' This has garnered a significant amount of investigation into the ethics of blame, and what are

sometimes referred to as 'standing conditions' on blame: conditions that would-be blamers need to meet for their blame to be unobjectionable. Plausible candidates on standing include that the matter be the blamer's business and that the blamer have sufficient evidence of culpable wrongdoing. But the condition that the blamer not be guilty of a similar fault as the target (in some sense to be specified) has received the most interest, and is perhaps the condition most familiar to our moral practices, given our many cultural injunctions against it (Cohen 2006).[1]

Given our desire for (1) and (2), it will be useful to compare two popular accounts of hypocritical blame and examine their contrastive strengths and weaknesses. Consider what I will call the *Moral Equality Account of Hypocritical Blame* from Fritz and Miller (2018, 2019).[2] They argue our right to blame is grounded in our acceptance of the moral equality of persons. Paradigmatic hypocrites, however, have an unfair differential blaming disposition. They are disposed to apply different standards to themselves than to others, thereby implicitly rejecting this equality and forfeiting their right to blame. This explanation seems plausible, scoring well on (2). However, the account has some flaws by (1). It has been noted that it seems to entail e.g. that citizens of Western nations can't blame *any* terrorists, for instance, unless they are equally disposed to blame all similarly bad terrorists to a similar degree (Todd 2019). This seems like quite a high bar of consistency that agents need to clear in order to retain their standing. Relatedly, it entails that we should object to merely inconsistent blamers for the same reason we object to hypocrites: since they apply different standards to others, they reject the

[1] If readers want more details on how I am thinking about 'standing' or 'legitimacy' or related notions, a strength of my argument is that little will turn on how one prefers to characterize such terms or their relevance.

[2] Another account which emphasises equality is Wallace's (2010), though it also faces objections regarding both (1) and (2) (Todd 2019; Fritz and Miller 2019).

equality of persons, so should thereby lose their standing to blame. But this doesn't seem to occur; while inconsistent blamers exhibit a certain fault (Telech and Tierney 2019; cf. Piovarchy and Siskind, forthcoming) our objections to merely inconsistent blamers seem to have different content to those directed at hypocritical blamers. When we object to inconsistent blame, we are exhorting our target to treat others fairly, which they can do *by* either increasing their level of blame towards one party, or decreasing their blame at the other. But when we object to hypocritical blame, we are typically objecting to their right to blame for this type of wrong at all: hypocrites "lack the standing to blame others *before* they blame" (Fritz & Miller 2022, p. 846). A final problem is that it counterintuitively entails that hypercrites—people who are disposed to blame themselves more than they blame others—also lack the standing to blame others (Fritz and Miller 2022; cf. Lippert-Rasmussen 2020; Tierney 2021).[3]

The other most popular account is what I'll call a *Commitment Account of Hypocritical Blame* (Friedman 2013; Piovarchy 2023a; 2023b; forthcoming; Rossi 2018; Isserow 2022). Todd (2019) argues that blame from agents is objectionable when they are insufficiently committed to the kinds of values that would condemn the wrong in question. Two kinds of cases act as particularly strong evidence that commitment matters, and that the account scores well by (1). First, agents can blame unobjectionably despite having committed a relevantly similar wrong if they are now committed to the relevant values. If the wrong was a long time ago and the agent has atoned and made reparations (which we expect people who are committed to relevant values to do), their blame will not be objectionable. Second, blame from agents is objectionable when they

[3] Additional problems for this account are discussed below in §5. Fritz and Miller point out that hypercrites seem like a problem for Todd's account too, but I'll show this can be dealt with.

lack sufficient commitment, even if they haven't carried out a relevantly similar wrong. There is something illegitimate about blame from someone who would do wrong given the chance, but who has simply not yet had the opportunity. Alternatively, one might be guilty of a different wrong that would be condemned by the same values. A benefit of this result is that it successfully explains why agents who are complicit in wrongdoing also often seem to lack the standing to blame (cf. Bell 2013): both complicity and paradigmatic hypocrisy manifest insufficient commitment to the relevant values. Someone who funds terrorism for profit might be very unwilling to detonate the bombs themselves, which may make accusations of hypocrisy inapt, but they are still insufficiently committed to the same values the terrorists are not committed to, and so they lack the right to blame said terrorists.

Unfortunately, while a *Commitment Account of Hypocritical Blame* does great by (1), delivering intuitive verdicts of when hypocritical blame is objectionable, it currently lacks even the beginning of an answer to (2). Proponents simply have not attempted to offer an explanation for why hypocritical blame is objectionable, with Todd (2019) suggesting that perhaps "there are no deeper moral facts from which [the account] can be derived" (p. 372). Finding an answer would make a significant contribution to our understanding of hypocrisy and appropriate blame.

While this literature on hypocritical blame has developed, another largely separate literature on the nature of blame has also been operating. A long-standing puzzle about blame's nature has been how to account for its wide variety of forms, which seem to resist analysis of some core, shared features. Blame seems to communicate something, but it can also be very private. It seems to involve reactive emotions, but can also be dispassionate. It seems to treat its target negatively, but sometimes people don't mind being blamed. A number of phenomena and aims are plausibly involved, such as beliefs,

attitudes, emotions, demands, marking of an impaired 'moral relationship', communicating a message, aligning moral understanding, and philosophers have used permutations of these factors to construct multiple theories, all of which face counterexamples (Shoemaker 2017; Shoemaker & Vargas 2021).

Recently, Shoemaker and Vargas (2021) have argued that what instances of blame all have in common is they are a costly and thus credible *signal*. Signalling has been used to explain a wide range of traits and behaviour in a variety of disciplines: bright colours on frogs signal toxicity, a peacock's tail signals good genes and access to resources (Zahavi & Zahavi, 1999), an engineering degree signals that one has a certain level of competence with engineering, and a real estate agent's expensive car signals that they have a lot of money (and are thus probably making high sales). What, exactly, is blame a signal of? As it turns out, an agent's *commitment* to norms and willingness to enforce them. The *Costly Signal Theory of Blame* and the *Commitment Account of Hypocritical Blame* were developed entirely independently, but seem practically made for one another. I believe this is not a coincidence. Together, they can help us provide (2), but it will take some work to see how.

Shoemaker and Vargas offer some brief comments on hypocrisy, mentioning that hypocritical blame seems to be "off" or somehow pointless (2021, p. 595). But their treatment is lacking in two ways. Firstly, their treatment doesn't amount to an explanation of why we *object* to hypocritical blame in the very particular ways that we do, rather than simply dismissing it. Second, and relatedly, their account is mostly descriptive in nature, not normative. They offer an account of what blame is, but not when it is appropriate, what justifies blaming, or why we are justified in objecting to certain forms of blame independent of that blame's fittingness and proportion. To answer (2), we need to somehow move from an account of what blame is, to when it is

legitimate. This is what I will provide. But to first understand what blame is, we need to understand what it does, and that requires understanding what problems it helps solve.[4]

3. Why Blame is Valuable

To understand the problems blame helps solve, we need to understand *mixed-motive public goods games*. In these 'games', agents must choose to put some resources either into a 'pot' (co-operate) which will then be redistributed to everyone, or keep their resources ('defect'). They face the following incentives. The money that goes into the pot experiences a multiplier before being redistributed. If everyone co-operates, everyone does much better, particularly because they can *reinvest* the extra benefits, making everyone do even better yet again. But, if one agent defects while the others co-operate, the former does fantastically as they get both their starting resources and some of the redistributed pot, while those who contributed end up worse off than if they'd never contributed at all. However, if everyone follows this strategy, then no-one contributes, and everyone ends up significantly worse off.

These games model many situations agents find themselves in when deciding whether to act morally (Curry, Mullins & Whitehouse 2019). Many moral norms provide important collective benefits when practiced at scale, over the long run. Individuals can do better by defecting, but if too many people follow these incentives, everyone ends up worse off. If I don't pay my taxes I can save for a new car, but if everyone thinks like this

[4] Shoemaker and Vargas argue that costly signals are good grounds for reputation, which brings "all sorts of goodies, including (at a minimum) a solution to the many everyday prisoner's-dilemma-type situations" (p. 587). However, the way these benefits come about are not explained in detail, nor are they well-known to the wider literatures on blame, standing, or hypocrisy, so it will help to investigate how.

there won't be usable roads to drive on. If I break my contract to build your house I can have a fun weekend spending your money, but if everyone does we'll be unable to count on anyone to provide us with services. Given these situations, most agents tend to act as conditional co-operators: they will co-operate if they believe that a sufficient number of others will too.[5] Such dispositions have been shown to be particularly competitive against other strategies (Axelrod & Hamilton, 1981), and have been used to model how moral behaviours evolved or are sustained by a number of theorists (Skyrms 2003; Bicchieri 2005; Brennan & Pettit 2005; Brennan et al. 2013).

Given our interest in maintaining co-operation, two things are incredibly valuable to have: assurance and punishment. Even if every agent individually wants to cooperate (conditional on others doing so), defections can still prevail because each agent may believe that the others won't cooperate. The mere worry that others will defect (or even knowledge that everyone's desire to cooperate isn't *common* knowledge) can be enough to make defections actually occur, in a self-fulfilling prophecy. If, however, there is some way for agents to assure one another that their cooperation will not be taken advantage of, then cooperation is much more likely to develop and sustain. Once I know you won't take advantage of me if I co-operate (because you're *committed* to co-operating), I'm going to be much more willing to co-operate and rely on you.

Some agents, however, prefer to defect even if they know that everyone else will cooperate. Since assurance of cooperation doesn't result in these agents cooperating, something else is needed. One particularly useful way of ensuring that such agents do

[5] What about cases where agents would continue doing the right thing even if no-one else did? This agent will have *internalized* the norm. Some individuals might be particularly virtuous in some domains, but dispositions to continue co-operating will be unlikely to persist or manifest if they keep making the agent worse off, particularly when we are interested in practices that take place at scale, over the long term. For more details see Piovarchy (ms).

not defect is by making defection costly, incentivizing cooperation. This is why punishment is useful: punishment increases the costs of defecting, thereby deterring it. This is particularly beneficial because staving off small numbers of initial defectors prevents a cascade of defections over iterated rounds as agents become continually less assured that others will co-operate (Fehr & Gachter 2000). If I would defect upon seeing three people defect, this will trigger anyone who would defect upon seeing four people defect, and so on. Stopping the initial defections from occurring greatly contributes to making cooperation stable. Of course, it's not enough to just have punishment, we need agents to believe there will be punishment, and so being assured of punishment is also useful, deterring agents from defecting in the first place. It is also valuable for current cooperators to be assured that if a defection occurs, it is unlikely to be followed by more defections from others.

We are now ready to see why blame is so valuable. Blame signals that I am committed to the norm (i.e. that I will follow it or co-operate) and that I will enforce the norm (not tolerate defectors, by punishing them, or at least, reducing the extent to which I assist them).[6] But what enables blame to play this signalling role (i.e. to distinguish me from someone who is not committed to the norm, but who would benefit from being able to make others think they are) is that it is costly: it is the kind of thing that requires incurring costs one would be unlikely to incur if one was not committed to

[6] Biologists and game theorists have particular characterisations of 'punishment' that may not fully overlap with the kinds of behaviours we are interested in. Some biologists define punishment as 'negative reciprocity', and some require that the behaviour be immediately costly. But as Jensen (2010) points out, there are various counterexamples to such proposals and a considerable diversity of punishment-like behaviour in animals to account for. These details do not matter for us. What does matter is that in blaming you I am motivated to treat you negatively in some way, even if this is only by being more likely to cease co-operation should continued defections occur.

the relevant norms or values.[7] Blaming motivates us to cease co-operation with the blamed and thus lose out on any benefits we were getting. The experience of blaming is generally unpleasant, and we also risk that the target will retaliate. The costliness makes it credible: if you didn't care about the norm violation, you probably wouldn't bother blaming. In contrast, if I regularly get outraged about theft, and refuse to interact with thieves, you can probably count on me to not steal your things much more than you can count on someone who appears indifferent towards theft.

This explanation of blame's nature gets most of the way to understanding why hypocritical blame is objectionable: we have a *very* strong interest in maintaining blame's ability to act as a credible signal of commitment. But hypocritical blame is *dishonest signalling*.[8] Hypocritical blamers are agents who signal their commitment to norms that they are, in fact, not sufficiently committed to (often evidenced by their past violations of said norms).[9] If we allow that kind of signal to proliferate unchecked,

---

[7] An anonymous reviewer at another journal objects that blame can deliver benefits to blamers too, and thus doesn't seem always costly. This misunderstands how costly signalling accounts work. The cost is calculated by comparing agents who have the relevant quality and agents who do not, not simply the costs to the agent with the quality. In our case, we are interested in the *differential* costs that would be incurred by someone who did not in fact care about the norm violation, compared to someone who does care about the norm. As Fraser (2012) explains, the threat of punishment for dishonest signalling can itself be part of what makes the honest signalling costly. This is why, even if it seems unlikely that blame could ever lose its signalling ability entirely (perhaps because of our inherent psychological makeup and evolutionary history), imposing costs on hypocritical blamers still helps blame maintain its reliability. Blame is a costly and thus reliable signal in part *because* we blame hypocritical blamers, and we are justified in objecting to hypocritical blame to help keep the signal reliable.

[8] Further evidence that dishonest signalling is the primary reason we dislike hypocrisy also comes from a number of psychological studies (Jordan, Sommers, Bloom, & Rand 2017). Note that even if it is common knowledge that a blamer is not committed, this doesn't thereby make their signalling fail to be dishonest, as an anonymous reviewer at another journal erroneously thought. The honesty is determined by whether the signal (blame) is sent by someone who has the relevant quality (commitment), not the blamer's sincerity or observer's knowledge of their qualities.

[9] For a more detailed discussion of how to assess commitment see Piovarchy (2023a; 2023b).

observers find it harder to distinguish between who can be counted on to comply with and enforce norms, and who cannot. While small levels of dishonest signalling can be tolerated, if it becomes too frequent the signalling system suffers as observers can no longer tell the difference between honest and dishonest signallers, eventually ignoring signals altogether (Mappes & Alatalo, 1997). This means we have a strong interest in combatting hypocritical blame. Objecting to hypocritical blame, disputing its legitimacy, calling attention to the mismatch between the agent's blame and their behavior, or blaming the hypocrite in turn, are all ways of doing this.

This, I propose, demonstrates that a *Commitment Account of Hypocritical Blame* is capable of answering (2), resolving its primary weakness and giving it a considerable lead over competing accounts of the badness of hypocritical blame. (As we shall see, there are more strengths to come). It not only provides intuitive answers about when blaming would be objectionable (as already noted), it does so while being perfectly consistent with a leading theory of blame that delivers intuitive verdicts about blame's diverse forms, and providing an extremely plausible, naturalistic explanation about our moral practices.

However, it would be even more satisfying to have a more detailed explanation of how this interest in calling out hypocritical blame connects up with an account of under what circumstances blame is objectionable. In particular, some readers may have questions about how we ought to respond to instances of blame that would not seem to threaten the signalling system very much. The next section shall take up this task.

4. From Valuable to Justified

We can understand how this macro-level story has implications for particular agents by introducing a *two-tiered* justification of our moral responsibility practices. These have

been developed by Rawls (1955), Dennett (with Caruso 2021), Barrett (2020), and most thoroughly by Vargas (2013).[10] Existing accounts have focused on making sense of the conditions that potential targets of blame must meet for blame to be appropriate (i.e. why they need to be a culpable wrongdoer), but as we shall see, this approach can also easily account for why potential blamers must also meet certain conditions.

The key move is to make a distinction between the external justification of a practice as a whole (in our case, what justifies a blaming system), and the conditions internal to that practice (under what circumstances it is appropriate for individual blamers to blame). While the former is forward-looking, being justified by the considerable benefits the practice provides, the latter are backward-looking, being concerned with what agents have previously done or what is true of them at the time of blaming. For instance, in asking whether it would be appropriate to blame this particular wrongdoer, we do not ask whether it seems that blaming here and now will serve consequentialist goals (in our case, stabilizing co-operation by providing assurance). We ask only whether that target is blameworthy, i.e. a culpable wrongdoer, which is determined by facts about the agent at the time of wrongdoing.[11] On a two-tiered account, what matters is that this *type* of wrong act is the kind of thing which, if blamed, would produce good consequences (over time, at scale).

An analogy might help. In basketball, there is a system of fouls which serves the goals of stopping players from gaining an unfair advantage and keeping to keep the game enjoyable for the audience. But everyone agrees that players who foul should be

[10] Such accounts have also been used in various guises by Piovarchy (2021), Milam (2021), Alfano (2021), Pereboom (2021), Roubichaud (2021), and Jefferson (2019).
[11] i.e., whether they had the capacity to avoid wrongdoing, in the case of Control theorists, or whether their actions expressed an objectionable evaluative attitude, in the case of Attributionists.

penalized *even when* that player did not gain an advantage and *even when* audiences would not enjoy the game more.

While this account is rule consequentialist in spirit, it avoids all of the objections rule consequentialism commonly encounters.[12] For example, it will not allow scapegoating. Blaming innocent people is clearly not the kind of thing that will, as a practice, incentivize co-operating. If I'm going to get blamed regardless of whether I co-operate or defect, what's the point of co-operating? Likewise, it does not allow letting wrongdoers off the hook whenever doing so seems like it will produce better consequences. It also successfully deters one-off violations, which are a known problem for consequentialist theories of punishment and game-theoretic strategies where punishment is engaged in with the intention of deterrence. Since potential wrongdoers know that consequentialist punishers wouldn't have reason to deter *after* the wrong, they reason that they could get away with a single wrong (defection), and thus commit it. However, if they know that blame and punishment attach to the type of act (e.g. murder) rather than the token act (this particular murder carried out by someone who will never murder again), they will be successfully deterred from even one-off norm violations. (Interestingly, this means that *not* aiming to deter can actually be *more* effective at deterring).

Now that we've seen how the account makes sense of the conditions agents must meet to be appropriate targets of blame, we can also see how the account easily makes

[12] The most thorough examination of how these accounts numerous avoid other commonly raised objections, which this paper endorses, can be found in Vargas (2013, p. 187–195). If these approaches strike readers as the wrong kind of reason, it is open to us to say that our moral practices need to be structured to work around forms of treatment that are inherently wrong. If scapegoating is inherently immoral, we can stipulate that our blaming practices must treat the wrongness of scapegoating as a fixed point, given first-order moral considerations are more fundamental.

sense of the conditions would-be blamers must meet for their blame to be appropriate too. For blam*ers* to be in a position to blame, we do not ask whether allowing *that* particular token act of blaming would produce good or bad consequences (namely, threatening the functioning or efficacy of the signaling system altogether). We instead ask whether this *type* of act is the kind of thing that would, at the level of a practice, produce good or bad consequences.

This, I propose, is how *The Commitment Account of Hypocritical Blame* provides a comprehensive answer to (2). Given the benefits that blame provides by being a credible signal, it straightforwardly follows that if that signal were to be undermined, we would lose those collective benefits. Since hypocritical blame is dishonest, if it is allowed to proliferate it can lead to the signaling system collapsing. Our interest in avoiding this makes it legitimate to object to token instances of hypocritical blame. To decide whether a particular hypocritical act is objectionable, we do not need to ask whether objecting now will produce the best consequences e.g. by looking at the number of agents who would see the signal, and their level of conditional cooperation, and the number of current defectors. We need only ask whether this agent is sufficiently committed to the values that they are trying to signal their commitment to. If they are not, we are within our rights to object to them.

5. Strengths

This account has a number of strengths, many of which have not yet been noticed as desirable criteria for theories of hypocritical blame. The first, and most notable, is that this account enables us to understand why we also object to hypocrisy that does *not*

involve blaming.[13] This is something most other accounts of hypocritical blame have not attempted to account for (cf. Isserow & Klein 2017; Piovarchy 2023b) and, more importantly, are unlikely to be able to provide. Any account which ties the badness of hypocritical blame too closely to the blaming (e.g. one's right *to blame*, as Fritz and Miller do) cannot then account for the badness of hypocrisy that, by stipulation, does not feature blaming. The *Commitment Account of Hypocritical Blame* is far more parsimonious on this front: hypocrites can dishonestly signal commitment in ways that do not involve blame, such as through their pronouncements, and we have an interest in objecting to these dishonest signals too.

Another important strength is that the account explains why certain forms of *cancellation* can somewhat reduce our objections towards hypocrites to some degree.[14] Blaming while acknowledging one has committed similar faults, for instance, typically tempers our objections. This is what we should expect: an acknowledgement makes clear that the agent is trying to cancel what their blame would ordinarily signal, namely

---

[13] Todd (2023) later revises his account slightly, suggesting that *The Commitment Account of Hypocritical Blame* is grounded in a broader *Be Better* norm on criticism generally: "One must: criticize x with respect to standard *s* only if one is better than x with respect to standard *s*." (p. 1158). I reject this modification for two reasons: (i) *Be Better* seems outcome luck-sensitive in a way that his original commitment account is not (Todd 2019, p. 363). If Dan is not, in fact, a better drawer than Lucy, but criticises her, when she says 'let's see you do better' he can silence her criticism by doing a better drawing even if this performance is fluky. (ii) I don't think the *Be Better* norm makes sense of many cases of amateurs criticising professionals. When soccer fans criticise players for missing a goal, they aren't criticising the player for 'passing off' said kick as the work of a professional. We don't expect professionals to make every goal, and the criticism can remain apt even if the player has already scored multiple goals and thus has already performed at a level commensurate with being a professional soccer player. Nevertheless, it seems to me that the fans' criticism of players *for missing the goal* remains apt (though Todd does spend some time trying to address this objection, and readers' intuitions may vary).

[14] An anonymous reviewer demurs, so readers' mileage may also vary. At least, the account can explain this result for those who share the intuition. Little would be lost for the overall plausibility of the account if it turned out that such acknowledgements do not reduce objectionableness; we can just say that their attempted cancellation fails.

that they can be counted on to co-operate with certain norms. The *Moral Equality* account has a harder time explaining this: given that acknowledging fault doesn't manifestly change e.g. whether the agent has an unfair differential blaming disposition (though self-blame might), it shouldn't have any effect on how much we object to their blame. An agent acknowledging their own faults should make no difference to the degree to which we blame them, but this doesn't seem to occur in practice.

The account also explains why we often object to hypocrites even on matters that are not immoral. Compare two pastors who blame others for having homosexual sex. Many people object that this blame is inappropriate, on the grounds that homosexual sex is not wrong. But many people object *even more* upon finding out that one of the pastors has homosexual sex themselves, and blame him *for being a hypocrite*. Though *Moral Equality* accounts may try to say that the pastor exhibits a differential blaming disposition, this doesn't explain the extra objection. Appealing to his putative forfeiting of a right doesn't make sense; since homosexual sex is not wrong, the pastor *already* lacks any right to blame others on this matter. The commitment account, in contrast, perfectly explains our reactions: our interest in having blame be a credible signal of commitment provides reasons to object to hypocritical blame generally, even in cases where blame is already unfitting.[15]

The account also explains why we sometimes have mixed feelings about whether some inconsistencies between walk and talk qualify as hypocritical. Suppose an avowed vegetarian eats meat on one occasion. If they were to blame regular meat-eaters, would

[15] Note this reason can be outweighed, such that one should not object here, all-things-considered, e.g. we might think that objecting in these cases would mistakenly be interpreted as communicating disapproval of homosexuality, or cruelly create pressure to hide one's sexuality. Even if objecting is inappropriate (Dover 2019), we need an explanation for why so many people *do* object, and why we often need arguments to the contrary.

they be hypocritical? Our mixed intuitions stem from the fact that we are unsure whether eating meat one time qualifies as being insufficiently committed to the values that condemn eating meat. On the one hand, it is a clear violation of said values. On the other hand, the vegetarian's success at abstaining from meat in many other instances counts as evidence that they are sufficiently committed. Insofar as 'commitment' is scalar, we should expect vague cases. 'Sufficient commitment' also need not be understood as a well-defined threshold (where all commitment below that point is insufficient, while all commitment above that point grants standing); we can allow there to be degrees of objectionableness despite talk of thresholds by modelling the relationship between 'degree of commitment' and 'legitimacy of blame' as following a sigmoid curve (cf. being 'tall').

The account can also deliver intuitive verdicts about hypercrites: agents who blame themselves more than they blame others. Fritz and Miller (2022) acknowledge such agents have a differential blaming disposition, but they make a partners-in-crime argument that such agents also pose a problem for accounts like Todd's. They propose that insofar as hypercrites necessarily fail to blame one party to the level that they ought, they do seem guilty of failing to take certain norms seriously, which we could cash out as failure to be sufficiently committed to the norm.

I think the case of the hypercrite is underdescribed. If the hypercrite reliably follows the norm themselves in most circumstances, this suggests they are committed to following the norm, and thus have standing to blame norm-violators, including themselves on the occasions where they violate the norm. If they never blame others, this might cause us to doubt whether they are committed to the norm's enforcement. But three things are worth noting. The first is that it plausibly matters *why* our hypercrite blames others less. Suppose the reason they fail to blame others is that they

have an admirable commitment to forgiveness and tolerance. Here, their failure to blame others may not be evidence of insufficient commitment to the norm or value, but simply evidence of commitment to another worthy value. The second is that hypercrites can still be disposed to enforce norms in other ways that don't involve blaming. If they are, failure to blame need not be good evidence of insufficient commitment. We can imagine Martin Luther King-types who are committed to forgiveness and reconciliation, but who still hold others responsible and make clear that they will not tolerate injustice. The third is that how often one follows norms is plausibly weighted much more heavily in our assessments of commitment than how often one enforces norms. The fact that 'hypercrite' is not a term in common usage (unlike 'hypocrite') is some evidence that we care much more about agents who treat themselves as exceptions to normative expectations than agents who let everyone else off.

Still, an opponent might press: what should we say about someone who blames themselves, but never blames others, and is never disposed to enforce norms, not out of commitment to some other value but simply because they don't care about those violations? I think it's hard to have any clear intuitions. But this is exactly what we should expect: because 'commitment' is made up of several components, we should expect to have unclear intuitions about agents who score very highly on some components (blaming themselves and following the norm) and very lowly on others (never enforcing the norm and never caring about others' norm following). Because such agents would be rather psychologically *abnormal*, I don't think such cases are a great guide for thinking about the structure of our moral practices. Even if we concede that such agents lack standing to blame and this is counterintuitive, that they make up a very small subset of all the cases of hypercrisy that originally motivated the worry means this is not especially costly for our overall account.

The account also easily handles a recent argument from Isserow (2022). Isserow argues that the *Commitment Account of Hypocritical Blame*, as standardly understood, delivers unintuitive verdicts about some subjunctive hypocrites, who would do the wrong thing given the chance due to insufficient commitment. In particular, she worries about how questions of standing to blame interact with cases of moral luck, such as with Nagel's (1979) German who moves to Argentina prior to WWII, but who would have become a Nazi had he stayed. Isserow argues that if the German in Argentina sees the Nazis on the news, and seethes with outrage at their actions, we should not object to this, even if we know that he would have become a Nazi had he stayed. She proposes that the reason he is entitled to blame is that he has not *manifested* insufficient commitment in any way.

There are two ways our account can accommodate this case, either of which I am happy to accept. The first is to question Isserow's handling of the case. Since what we are interested in is degree of commitment, it would help to have some idea of how quickly the man would have become a Nazi, and this detail is left out of the thought experiment. If the man would have willingly become a Nazi merely one day after missing his flight to Argentina, then I would say that he does lack sufficient commitment, and is sorely mistaken about his values when he is sitting in front of the television one day after arriving in Argentina. Once we know this detail, it is far more plausible that we would find something criticisable about his outrage, especially if we could peer into a nearby possible world and see him donning a swastika. Far more likely, however, is that when we say he would have become a Nazi had he stayed, we mean that he would have *eventually* become a Nazi, over time. But in such a case it is much more plausible to say that by the time he became a Nazi, his commitments had, in fact, changed. The man who has been in Argentina for some time, however, has not had

his commitments change, and thus nothing is objectionable about his outrage. His response is an accurate and honest signal of his present commitments.[16]

The alternative approach is to modify our *Commitment Account of Hypocritical Blame* slightly. Rather than emphasising actual level of present commitment, it could instead emphasise level of manifested commitment, without significantly changing the intuitive appeal of Todd's account. (Since most hypocrites *have* previously manifested insufficient commitment, the verdicts it gives about key thought experiments will remain largely the same, save for some subjunctive hypocrites). When someone finances terrorism, murders an innocent, or cheats on their spouse, that agent is thereby manifesting insufficient commitment to the kinds of values that would condemn such actions.

A consequence of this change is that even if someone became sufficiently committed to the relevant values after violating them, they would not regain their standing to blame until they had adequately manifested or signalled that commitment. Suppose someone cheats on their spouse, realizes how wrong this was, is guilt-ridden, and the

[16] Isserow stipulates that she is considering this to be a case of synchronic circumstantial moral luck, such that the German is presently susceptible to influence of propaganda and corrupt authorities, and thus tries to ward off this response. But 'susceptibility' is underspecified, and the way she ends up describing him strongly suggests that he would not become a Nazi immediately, and that the forces that make him into one would do so *by* changing his commitments. This makes the details of the case somewhat inconsistent and thus insufficient for adjudicating whether actual or manifested commitment is what matters. She is right that "It seems a stretch to maintain that this man—who, we may imagine, has never believed in Aryan superiority, nor betrayed any hint of ill-will towards Jewish people (but easily would have, had he stayed)—lacks the entitlement to blame his former compatriots for their moral crimes" (p. 182). But it equally seems a stretch to say that he is either presently *not* committed to values which condemn Nazism, or would culpably commit Nazi acts *without* his commitments first changing. As a result, it does not follow from my earlier (2023a) argument that most ordinary people lack the standing to blame the subjects in Milgram's (1963) obedience to authority experiments (since most people would have behaved similarly) that most people thereby lack the standing to blame Nazis.

very next day is extremely committed to fidelity. Despite their actual level of commitment, we might think that they don't regain their standing to blame other philanderers until they have adequately manifested that commitment; say, by being faithful for an extended period of time, showing remorse, atoning, and repairing the harm that they have caused (Radzik 2009; Piovarchy 2020b). This result seems very plausible.

The key difference between these two responses concerns whether we think standing to blame ought to depend on level of actual commitment, or merely manifested commitment. While there may be some principled reason to favour one of these rather than the other, neither approach currently seems to face strong objections, nor do they jointly lead to any dilemma for the account, so I am happy to leave this detail aside for others to develop.

Another strength of the account is that our explanation has been achieved without relying on any contentious analysis of what standing is, how to best understand the exact relation that obtains between hypocrites and their blame. Elsewhere I've (2020a) argued that it is often unclear what exactly 'standing' amounts to, noting that there are problems with describing what standing amounts to in terms of fittingness, appropriateness (Isserow 2022), rights (Fritz and Miller 2018), legitimacy (Duff 2010), permissible deflection of reasons (Herstein 2017), or a general ethics of blame (Rossi 2020), which have variously been employed by philosophers to frame their investigations into hypocrisy.[17] Though I have chosen to frame our investigation primarily in terms of 'objectionableness' to be as neutral as possible on these matters

[17] There I argued that the most plausible interpretation concerns normative powers (given this also makes sense of standing to forgive, with this insight then later being adopted by Fritz and Miller 2022). I am not yet sure how to square the account in that paper with this one.

(anchoring our investigation by reference to our reactions to hypocrites), our explanation—that we have a strong interest in maintaining blame's signalling function—does not rely on any particular understanding of standing, or any particular analysis of any of these related terms. Our answer, then, is quite ecumenical regarding competing conceptions of these issues.[18]

Finally, this account also preserves the intuition that hypocrisy seems to have some connection with dishonesty or false pretense. Several earlier accounts of the nature of hypocrisy characterize hypocrites as deceiving others in some way. McKinnon (2002) argues that hypocrites are overly preoccupied with their moral reputation or a desire to be thought well of, and this leads them to misrepresent their motivations. Szabados and Soifer (2004) characterise hypocrites as people who pretend their actions are motivated by moral reasons when they are not. Kittay (1982) thinks hypocrites misrepresent themselves as conforming to norms in important domains where sincerity matter (e.g. morality, friendship). These formulations, however, all face objections (Isserow and Klein 2017; Rossi 2020), namely that they fail to distinguish hypocrisy from other forms of deception, or tie hypocrisy too closely to criticisable motives that need not be present to count as hypocritical (as when a parent hides their smoking from their children, while extolling the importance of avoiding cigarettes). Our account avoids these objections: other forms of deception (e.g. lying) are not centered around signalling of commitments, and what one signals in their actions (and whether it is criticisable) need not depend upon motives. But our account also retains much of the original appeal of these accounts. Signalling behaviors contribute to one's moral

[18] Though in this paper I occasionally spoke of appropriateness or legitimacy for consistency with others, nothing would be lost by formulating our explanation simply in terms of when we should or should not object to blame.

reputation (McKinnon's target), hypocrites do misrepresent themselves as being committed to norms (and this is very close to conforming with norms, which Kittay emphasized), and agents who are motivated by moral reasons (Szabados and Soifer's concern) are typically committed to the values that would approve of those reasons. Without first understanding the nature of signaling, or the ways that norms can (fail to) be sustained, it is hard to articulate the exact way that hypocrites are dishonest, but once these are on the table, it becomes clear why many previous accounts all circled around a common theme.

6. Conclusion

The Commitment Account of Hypocritical Blame, and *The Costly Signalling Theory of Blame*, are each notable contributions to our understanding of our moral responsibility practices. With some work, they can together help us understand not only why we do object to hypocritical blame, but why we ought to. Hypocritical blame is dishonest signalling, and we have an interest in keeping it in check to maintain the benefits our current blaming practices provide.

Acknowledgements

This paper benefitted from my attendance at The Honesty Project Summer Seminar at Wake Forest University, which was supported by The John Templeton Foundation.

Funding Information

None

ORCID

Adam Piovarchy 0000-0002-5169-2030

References

Alfano, Mark (2021) 'Towards a Genealogy of Forward-Looking Responsibility', *The Monist* **104**: 498-509. doi:10.1093/monist/onab015.

Barrett, Jacob (2020) 'Optimism about Moral Responsibility' *Philosophers' Imprint* **20**(33): 1-17.

Bicchieri, Cristina (2005) *The Grammar of Society: The Nature and Dynamics of Social Norms*: Cambridge University Press.

Brennan, Geoffrey, & Philip Pettit (2005) *The Economy of Esteem: An Essay on Civil and Political Society*: Oxford University Press.

Brennan, Geoffrey, Lina Eriksson, Robert E Goodin & Nick Southwood (2013) *Explaining Norms*. Oxford: Oxford University Press UK.

Curry, Oliver S, Daniel A Mullins, & Harvey Whitehouse (2019) 'Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies', *Current Anthropology*, **60**: 47-69. doi:10.1086/701478.

Dover, Daniela (2019) 'The Walk and the Talk', *Philosophical Review* **128:** 387-422. doi:10.1215/00318108-7697850.

Fehr, Ernst, & Simon Gachter (2000) 'Cooperation and punishment in public goods experiments', *American Economic Review* **90**: 980-994.

Fraser, Ben (2012) 'Costly signalling theories: beyond the handicap principle', *Biology & Philosophy* **27**: 263-278. doi:10.1007/s10539-011-9297-8.

Friedman, Marilyn (2013) 'How to Blame People Responsibly', *Journal of Value Inquiry* **47**: 271–284. doi:10.1007/s10790-013-9377-x.

Fritz, Kyle G, & Daniel Miller (2018) 'Hypocrisy and the Standing to Blame', *Pacific Philosophical Quarterly* **99**: 118–139. doi:10.1111/papq.12104.

Fritz, Kyle G, & Daniel Miller (2019) 'When Hypocrisy Undermines the Standing to Blame: a Response to Rossi', *Ethical Theory and Moral Practice* **22**: 379–384. doi:10.1007/s10677-019-09997-3.

Fritz, Kyle G, & Daniel Miller (2022) 'Two Problems of Self-Blame for Accounts of Moral Standing', *Ergo* **8**: 833–856. doi:10.3998/ergo.2255.

Herstein, Ori (2017) 'Understanding standing: permission to deflect reasons', *Philosophical Studies* **174**: 3109–3132. doi:10.1007/s11098-016-0849-2.

Isserow, Jessica (2022) 'Subjunctive Hypocrisy', *Ergo* **9**: 172-199. doi:10.3998/ergo.2263.

Isserow, Jessica, & Colin Klein (2017) 'Hypocrisy and Moral Authority', *The Journal of Ethics and Social Philosophy* **12**: 191–222. doi:10.26556/jesp.v12i2.224.

Jefferson, Anneli. (2019) 'Instrumentalism about Moral Responsibility Revisited', *Philosophical Quarterly* **69:** 555-573. doi:10.1093/pq/pqy062.

Jordan, Jillian J., Roseanna Sommers, Paul Bloom, & David G. Rand (2017) 'Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling', *Psychological Science* **28**: 356-368. doi:10.1177/0956797616685771.

Herstein, Ori (2017) 'Understanding standing: permission to deflect reasons', *Philosophical Studies* **174**: 3109–3132. doi:10.1007/s11098-016-0849-2.

Kittay, Eva F (1982) 'On hypocrisy', *Metaphilosophy* **13**: 277-289. doi:10.1111/j.1467-9973.1982.tb00685.x.

Lippert-Rasmussen, Kasper (2021) 'Why the moral equality account of the hypocrite's lack of standing to blame fails', *Analysis* **80**: 666-674. doi:10.1007/s10892-023-09454-5.

Mappes, Johanna, & Rauno V Alatalo (1997) 'Batesian mimicry and signal accuracy', *Evolution* **51**: 2050-2053. doi:10.2307/2411028.

Milam, Per-Erik (2021) 'Get Smart: Outcomes, Influence, and Responsibility', *The Monist* **104**: 443-457. doi:10.1093/monist/onab011.

Milgram, Stanley (1963) 'Behavioral study of obedience', *The Journal of abnormal and social psychology* **67**: 371. doi:10.1037/h0040525.

Nagel, Thomas (1979) *Mortal Questions*. New York: Cambridge University Press.

Pereboom, Derk (2021) 'Undivided Forward-Looking Moral Responsibility', *The Monist* **104**: 484-497. doi:10.1093/monist/onab014.

Piovarchy, Adam (manuscript) 'Signalling, Sanctioning and Sensitising: How to Uphold Norms with Blame'.

Piovarchy, Adam (forthcoming) 'Epistemic Hypocrisy and Standing to Blame', *Erkenntnis*.

Piovarchy, Adam (2023a) 'Situationism, subjunctive hypocrisy and standing to blame', *Inquiry: An Interdisciplinary Journal of Philosophy* **66**: 514-538. doi:10.1080/0020174X.2020.1712233.

Piovarchy, Adam (2023b) 'Does Being 'Bad Feminist' Make Me a Hypocrite? Politics, Commitments and Moral Consistency', *Philosophical Studies* **180**: 3467-3488. doi:10.1007/s11098-023-02056-9.

Piovarchy, Adam (2020a) 'Hypocrisy, Standing to Blame and Second-Personal Authority', *Pacific Philosophical Quarterly* **101**: 603-627. doi:10.1111/papq.12318

Piovarchy, Adam (2020b) 'Blame in the Aftermath of Excused Wrongdoing', *Public Affairs Quarterly* **34**: 142-168. doi:10.2307/26921124.

Piovarchy, Adam & Scott Siskind (forthcoming) 'A Modest Defence of Somewhat Selective Outrage', *Ergo*.

Rawls, John. (1955). 'Two concepts of rules', *Philosophical Review* **64**: 3-32.

Riedener, Stefan (2019) 'The Standing To Blame, or Why Moral Disapproval Is What It Is', *Dialectica* **73**: 183-210. doi:10.1111/1746-8361.12262.

Roadevin, Cristina (2018) 'Hypocritical Blame, Fairness, and Standing', *Metaphilosophy* **49**: 137-152. doi:10.1111/meta.12281.

Rossi, Benjamin (2018) 'The Commitment Account of Hypocrisy', *Ethical Theory and Moral Practice* **21**: 553–567. 0.1007/s10677-018-9917-3.

Rossi, Benjamin (2020) 'Hypocrisy is Vicious, Value-Expressing Inconsistency', *The Journal of Ethics* **25:** 57-80. doi:10.1007/s10892-020-09340-4.

Robichaud, Philip. (2021). 'Characterizing the Value of Morally Responsible Agency' *The Monist* **104**: 458-470. doi:10.1093/monist/onab012.

Shoemaker, David, & Manuel R Vargas, (2021) 'Moral torch fishing: A signaling theory of blame' *Noûs* **55:** 581-602. doi:10.1111/nous.12316.

Skyrms, Brian (2003) *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

Szabados, Béla, and Eldon Soifer (2004) *Hypocrisy: Ethical Investigations*. New York: Broadview Press.

Telech, Daniel & Hannah Tierney (2019) 'The Comparative Nonarbitrariness Norm of Blame', *Journal of Ethics and Social Philosophy* **16**: 25-43. doi:10.26556/jesp.v16i1.654.

Tierney, Hannah (2021) 'Hypercrisy and Standing to Self-Blame', *Analysis,* **81**: 262-269. doi:10.1093/analys/anaa074.

Todd, Patrick (2019) 'A Unified Account of the Moral Standing to Blame', *Noûs,* **53:** 347–374. doi:10.1111/nous.12215.

Todd, Patrick (2023) 'Let's See You Do Better', *Ergo* **10:** 1157–1186. 10.3998/ergo.5178.

Vargas, Manuel R (2013) *Building Better Beings: A Theory of Moral Responsibility*. New York: Oxford University Press.

Wallace, R. Jay (2010) 'Hypocrisy, Moral Address, and the Equal Standing of Persons', *Philosophy and Public Affairs, 38*(4), 307–341. doi:10.1111/j.1088-4963.2010.01195.x.

Zahavi, Amotz, & Avishag Zahavi (1999) *The handicap principle: A missing piece of Darwin's puzzle*: Oxford University Press.