Synthese (2026) 207:135
https://doi.org/10.1007/s11229-026-05444-w

ORIGINAL RESEARCH

Signalling, sanctioning and sensitising: How to uphold norms with blame

Adam Piovarchy[1]

Received: 18 February 2025 / Accepted: 5 January 2026

Abstract

This paper provides a unified account of the nature of blame by taking a broader look at the connection between individual blaming reactions and the moral practices of communities. The methodological proposal is that to understand what blame is, we need to understand what it does, but to understand what it does, we need to understand what problems it helps solve. This, in turn, requires looking at the kinds of problems that communities have *qua* communities, namely, developing agents who are competent norm-followers and who are motivated to exercise this competence. The reason blame seems so heterogeneous is that it has not one, but several, connected effects: it signals, sanctions and sensitises. Signalling provides assurance that norms will be upheld and enforced. Sanctioning deters norm violations. Sensitising produces agents who have internalised the norms and who are skilled at complying with them. These effects combine to produce agents who are robustly disposed to uphold norms, which is a very valuable thing to create. This account not only solves a number of puzzles about the nature of blame, it does so while providing a simple and principled story, one which emphasises our own nature as norm-sensitive beings.

Keywords Blame · Moral responsibility · Reactive attitudes · Norms · Accountability

✉ Adam Piovarchy
adam.piovarchy@nd.edu.au

[1] Institute for Ethics and Society, The University of Notre Dame, Broadway, Sydney, NSW 2007, Australia

Published online: 23 March 2026

1 Introduction

Consider the following questions:

1. What is blame?
2. Under what circumstances is it appropriate to blame?
3. What justifies blaming in those circumstances?

Though this paper's interest is in (1), much of the existing literature on (1) grew while trying to simultaneously answer (2) and (3) too. It is easy to see the relevance: whether blame is ever justified depends on what kind of thing it is, and the circumstances in which we actually blame (and think it appropriate to do so) help us home in on the phenomenon under investigation. But historically, the reason for this connection is that blame is commonly taken to have some point, or to play some kind of role, in our *moral practices*, namely having some connection to the way we hold people responsible or accountable.[1]

It seems universally accepted that (3) requires an answer; no-one thinks blame doesn't matter. Finding justification for blame is often taken to be more important than finding justification for other emotional reactions or attitudes; indiscriminately hoping or being amused will get you some funny looks, but blaming indiscriminately will get you, well, blamed. Blame needs justification because—and as a partial answer to (1)—unlike many other emotions, it is commonly taken to have a "depth, force, or sting" (Hieronymi, 2004, p. 117), or "quality of opprobrium" (Wallace, 1994, p. 80).

The interest in accounting for blame's sting, and the connections between (1), (2) and (3), can both be seen in existing accounts of blame. Hieronymi (2004), for instance, thinks the force of blame stems from it being a set of judgements marking the significance of the target's conduct, namely as having impaired relations of mutual regard, and that when such a judgment is accurate, it cannot be considered unfair. This leads her to reject claims that blame requires a certain kind of control to be justified. Pereboom (2014, p. 180), in contrast, thinks that blaming attitudes like resentment are harmful and typically sustained by beliefs about free will. Since he thinks we lack free will, he thinks such forms of blame are unjustified.

It is notable, however, that contemporary accounts of blame's nature don't actually provide much reference to our 'moral practices,' the moral community, or blame's role in them. The moral community has been taken to play an important role in setting what the standards of conduct are, what we can reasonably demand of each other, or the facts about when agents count as blameworthy. But once these are set, they don't seem to be thought crucial to understanding blame's *nature*. Having identified that *that* thing—what a blamer directs at a target—is what we want to understand, the existing literature on blame almost universally then takes *that* thing as its primary focal point when developing theories of its nature. The norms we blame people for

[1] Perhaps being *the* way that we hold people accountable, with many philosophers notably taking there to be a distinct category of 'accountability blame' (Watson, 1996).

violating are treated as a given, and the community is implicitly treated as simply the individual blamer, multiplied.[2]

Absent from this literature is interest in the ways in which wrongdoing can propagate, or the distinct difficulties that communities face *qua communities*. Little attention is given to how communities work over time, or how they are distinct from mere collections of moral agents, and such factors are hard to notice unless one 'zooms out' from individual blamers. Isn't it plausible our moral responsibility practices might have been shaped by, or in response to, such factors? And wouldn't it thus be somewhat misguided to investigate blame's role in said practices while leaving those factors out of the picture? This thought is particularly forceful when one notes how little attention the moral responsibility literature gives to research on the nature of *norms*. Since both fields purport to be investigating significant swathes of our moral practices, it is rather surprising that they have hitherto been functioning as independent silos.[3]

Focusing on blame as a product of blamers—while setting aside questions of how our normative standards come about—can seem cleaner, cutting away baggage that otherwise makes a concept or subject matter difficult to pin down. But any promise of an increased handle on our explanandum has been hard to square with the now common observation that blame seems significantly 'disunified': it has a wide variety of associated phenomena, and it is very difficult to say how they are all connected, or why they all count as blame (Fricker, 2016; Sher, 2005; Shoemaker, 2017). In my view, this perceived heterogeneity is a direct result of the above narrowing of theoretical focus. It's not that blame doesn't paradigmatically involve an individual blamer and a target. But we end up with a myopic view of blame's nature when philosophers try to understand it by focusing on the immediate site of blaming reactions, or what helps individuals feel like their blame has achieved its point.

This paper provides a comprehensive account of blame's nature. The methodological proposal of this paper is that to understand what blame *is*, we need to understand what it *does*. But to understand what it *does*, we need to understand what problems it helps *solve*. And these are problems that will remain unnoticed if philosophers focus primarily on blamer-target dyads, forgetting that moral practices are upheld

[2] For example, Hieronymi (2004) and Scanlon's (2008) accounts of blame focus on an impaired relationship of mutual regard between the wrongdoer and individual blamers. Smith's (2013) account of blame as moral protest takes blame to seek acknowledgement from the wrongdoer and community, but it largely focuses on what the individual blamer is trying to achieve. Sher (2005, Ch. 1) offers some reflections on how alien life without blame would be, but his account ends up focusing on the blamer's belief-desire pair, associated dispositions, and commitment to morality. The reactive attitudes, which have received the lion's share of attention, are, of course, things that individuals experience e.g. Strabbing (2019) holds that "an agent is blameworthy for an action if and only if a negative reactive attitude is appropriate toward her on account of it" (p. 3122). Thinking of blame as an emotion (e.g. Portmore, 2022) also similarly emphasises its manifestation in the blamer.

[3] Brennan et al.'s (2013) *Explaining Norms* only uses the word 'blame' in one cursory paragraph regarding mental states, and Brennan and Pettit (2005) *The Economy of Esteem* fares no better. The term also appears only three times in Bicchieri's (2016) *Norms in the Wild*, and four times in *The Grammar of Society* (Bicchieri, 2005), mostly offhandedly. A full survey of engagement in the other direction is infeasible, but as some indication, *none* of these influential works are cited at all in *Blame: Its Nature and Norms* (Coates & Tognazzini, 2013) nor *The Oxford Handbook of Moral Responsibility* (Nelkin & Pereboom, 2022).

by *communities*, groups of people who need to live together and who can encounter very particular problems requiring very particular interventions. When we put blame into context and piece together several threads from existing accounts, an account of blame's nature emerges in which it has not one, but multiple effects. In a slogan: signalling, sanctioning and sensitising. These are united by common function: upholding norm-following.

Though making these claims is simple enough, I believe that the tendency to think about blamer-blamed dyads distorts philosophers' understanding of them. Consequently, ensuring that participants to this debate understand the account requires quite a bit of conceptual apparatus. This in turn requires a willingness from readers to take several detours into topics somewhat afield—public goods games, costly signalling, assurance, deterrence, norms, norm stability, norm-psychologies—before the account fully comes into view. While this may seem unnecessarily detailed to readers familiar with such things, many philosophers writing on blame's nature are insufficiently familiar with them and how they relate to one another, leading to frequent misunderstandings regarding not only the account's strengths, but what is even being claimed.[4] Part of this paper's goal is thus to present this apparatus at a level of detail that can help put to bed several common misunderstandings, explain why multiple common objections that have arisen are, in fact, no objection at all, and act as a common resource for future work engaging with such concepts. Once all this is on the table, we will have a simple, unified account of blame.

2 Cooperation

I have suggested that existing treatments of blame give insufficient attention to broader communities and the distinct problems that they face *qua* communities. To understand what I mean, we need to take a brief detour and consider typical *mixed-motive public goods games*, which many philosophers and scientists use as models of how moral behaviours and norms evolved (Bicchieri, 2005; Curry et al., 2019; Skyrms, 2003). In these 'games', individual agents must choose to either put some resource into a 'pot' (co-operate) which then be redistributed to everyone, or keep said resource ('defect').[5] They face the following incentives. 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 returns, making everyone do even better yet again. But, if one agent defects while the others co-operate, that agent does fantastically as they get both their starting resources and some of the redistributed pot, while those who contributed end up worse than if they'd never contributed at all. However, if everyone follows this thinking, no-one contributes, and no-one gets any benefit.

Co-operating, then, can be a very valuable thing for communities to have. What do we need to ensure it is maintained? My proposal is that in these settings, we want agents to have two distinct features:

[4] Evidenced also by reports from *several* reviewers of this very paper at other journals.

[5] The differences between this set up and e.g. the stag hunt do not matter for our purposes.

1) *Competence* regarding how and when to cooperate.
2) *Motivation* to exercise this competence.

Of course, agents can be motivated to cooperate to different degrees. Three tiers of motivation are particularly worth identifying:

2a) Motivation to co-operate so long as there are not more direct benefits to defecting.
2b) Motivation to cooperate so long as this does not come with a significant risk of being taken advantage of by other agents.
2c) Motivation to cooperate to such a high degree that one will do so even when it is costly or there is a significant risk of others defecting.[6]

Competence and motivation are clearly both necessary: communities cannot function well with agents who want to cooperate but don't know how, nor agents who know how to cooperate but don't give a damn about doing so. But a few factors relevant to people's level of motivation to cooperate are worth disentangling. Most people (depending on the game's set up) tend to approximate being *conditional co-operators*: while they want to co-operate, they don't want to be taken advantage of, so they'll only co-operate if a certain number of others are co-operating too. Strategies like this have been proven to be particularly effective against a large range of other strategies in simulations (Axelrod & Hamilton, 1981).[7]

What is useful among conditional co-operators, then, is if we can establish some kind of expectation that others will cooperate, which will make cooperation stable. 'Expectation' is used here in both a predictive and a normative sense: predictive, in that one believes others are likely to cooperate; normative, in that one feels everyone ought to cooperate. What is useful is establishing a *norm* of cooperating.[8]

The literature on norms thus seems like a natural ally in our investigation. But part of the reason leading accounts of norms have had so little engagement with the moral responsibility literature is they often make a sharp distinction between moral and social norms, and are only interested in the latter. Moral norms, they think, tend not to have the conditional nature of social norms.[9] For instance, while we are likely to

[6] How to precisify these tiers does not matter; this is just to illustrate the roles that assurance, sanctions and internalisation play maintaining norm-following.

[7] Famously known as variants of 'tit-for-tat'. Though in some games such strategies can be outcompeted by other strategies (e.g. 'Pavlov', which continues to defect when the other agent cooperates without retaliating), these can then be outcompeted by strategies that tit-for-tat outcompetes once again (Nowak & Sigmund, 1993).

[8] Though of course agents can also 'cooperate' to produce very unjust states of affairs. To be clear, I am not saying that all norm-following constitutes a form of cooperation within these kinds of settings, but such games act as useful models of the distinct ways valuable norms can fail to be upheld.

[9] There is some debate about how to best characterise what norms are. Some argue that they are patterns of behaviour in an equilibrium (Young, 1998) whereas others take them to involve features of agents. Bicchieri (2005, 2016), for instance, takes norms to be rules that agents have conditional preferences for complying with: they will comply conditional on other agents following the norm and believing that the norm ought to be followed. Brennan et al. (2013) take norms to be clusters of normative attitudes (roughly, a normative expectation) in a group, combined with knowledge that such a cluster of attitudes exists.

cite other people's behaviours or expectations in our justification for complying with the social norm that one should wear black at a funeral, such considerations won't be cited as justification for the norm of abstaining from murder. In contrast, theories of moral responsibility and blame are interested in both, focusing on wrongdoing and the 'demands' agent make of one another.

I believe that the moral/social distinction is a red herring for our purposes, for several reasons. Attempts to make the moral/social norm distinction on the basis of conditionality or social practices face many counterexamples, and whether such a distinction can be maintained at all is very questionable (Stich, 2018; Sugden, 2010). More importantly, our interest is in the *stability* of valuable norm-following, not how agents would justify their norm-following.[10] As even Bicchieri (2005, p. 20, fn. 10) notes, it *is* likely that if many others began murdering, you *would* become much more likely to murder yourself, and less disposed to sanction other murderers.[11] What matters for us is that even if some individuals could refrain from murdering regardless of how others were behaving, collective commitments to norms against murdering can be more or less stable over time, and in much the same way that merely 'social' norms are.[12] I thus propose that the literature on norms *is* a natural ally for thinking about the stability of moral behaviour within communities.

So, given we want agents to develop the motivation and competence required to comply with valuable norms, our question now is: how do we produce this? Here's how: we make agents competent (knowledgeable and skilled) by giving them continued *information* and *feedback*. We remove agents' incentives to defect by deterring defections with *sanctions*. We remove agents' wariness about being taken advantage of by providing them with *assurance*. And we produce agents who are motivated to a high degree with *internalisation*.

Each of these components requires unpacking. But the key idea here is that blame does *all* of these things: it signals, sanctions, and sensitises, which the remainder of this paper will elaborate on. Having multiple effects can seem *prima facie* theoretically unwieldy, but it will also be shown that the underlying story is actually quite simple, and our account requires very few moving parts in order for these effects to result.

3 Signalling

Suppose you're in the kind of situation modelled by a mixed-motive public goods game. You recognise the benefits that can be collectively gained if everyone cooperates. In fact, you want to cooperate—you're a good person! But you worry that if you cooperate while everyone else doesn't, you'll get taken advantage of and be left

[10] Additional reasons the connection between these two literatures have not previously been noticed (regarding the relevance of sanctions) are examined below (fn. 31).

[11] Thrasher and Handfield (2018), for instance, demonstrate that honour killings can be understood within a system of moral norms trapped at inefficient equilibria.

[12] Even if some moral behaviours or attitudes are innate, they can just be modelled as behaviours which provide other kinds of 'cooperative' benefits over longer timescales (Curry et al., 2019).

worse off. To work out if other players will cooperate with you, you need information. You could ask the players what they'll do. But this strategy has an obvious pitfall: the kind of person who would defect might be *dishonest* about their intentions, and lie to take advantage of you.

This kind of setting is very common in both human and animal environments. We have *senders* who would benefit from being able to let *observers* know they possess some hard-to-observe quality. Many animals would benefit from letting other animals know they would win in a fight, since this would enable them to avoid the costs of fighting, but relative fighting ability is not the kind of thing others can easily detect prior to said fight. Similarly, competent job applicants would benefit from letting employers know they are competent and reliable, but it may be hard for employers to identify this without employing applicants for several months.

To get around this problem, beings with the relevant qualities engage in *signalling*: they provide more easily observable information about themselves which communicates they possess the relevant quality. But a new problem is that signals can be exploited by beings which do not possess the relevant hard-to-observe property, deceiving observers. If poisonous animals can signal toxicity to potential predators through developing bright colours, non-poisonous animals can avoid predators by becoming brightly coloured too (Mappes & Alatalo, 1997). This 'dishonest' signalling is bad for observers and 'honest' signallers too. If a signal becomes unreliable, observers then ignore signals sent by everyone. This results in high-quality mates missing out on mating, strong fighters being challenged to fights, and competent workers not being hired.

Senders with the relevant qualities overcome this problem by developing signals that are hard-to-fake. One important method is to adopt signals that are *costly*: signals which require costs that organisms without the relevant quality would be unlikely to incur (because they typically outweigh the benefits that organisms without the quality would gain by signalling; Bird & Smith, 2005).[13] Growing a large tail requires lots of resources and high fitness, so peacocks with large tails are signalling access to resources and high fitness (Zahavi & Zahavi, 1999). Gazelles can engage in stotting: jumping up and down which expends energy that would be much better spent outrunning a lion, and which only a gazelle with large amounts of energy to expend would risk spending (Fitzgibbon & Fanshawe, 1988). Because talk is cheap, people who know lots about engineering signal their skills and education not by simply saying they know lots about engineering, but by obtaining an engineering degree. This is a much better signal than words and arguments because it is the kind of thing which is hard for people who lack the relevant skills or education to obtain.

Costly signalling has been able to explain a diverse range of phenomena across many disciplines, including evolutionary biology, social psychology, anthropology, and decision theory. Let's now examine how it partially helps us understand blame. Shoemaker and Vargas (2021) argue that blame signals lots of information about the blamer, namely, that the norm really matters to them and they are not willing to toler-

[13] A technical point which is frequently misunderstood: beings with the quality need not actually incur any costs; it is sufficient that enough beings without the quality *would* incur a cost. Costly signalling can be entirely costless in practice for beings with the relevant quality.

ate violations of it. People who are committed to a norm typically believe they and others should abide by that norm, feel motivated to abide by that norm, and experience a range of characteristic emotional responses towards other people's compliance or violations. They also typically abide by that norm when it would be reasonable to expect them to do so, and abstain from behaviours which would encourage others to violate the norm.

Identifying who has a genuine commitment to a norm can be difficult. Someone with these properties could just assert that they take a norm seriously, but such assertions can also be made convincingly by agents who are not committed to those norms. Reliably following the norm yourself would be good evidence one is committed, but there may be few opportunities to visibly follow some norms in ways that others can readily observe. For example, it's hard to definitively prove that one has never cheated on a past partner, and won't do so in the future with their present partner. Relying on previous behaviour to make an inference about commitments will also fail to screen out agents who have been following a norm for selfish reasons. Blaming, in contrast, is the kind of thing that's hard to fake and sustain, even for professional actors, and so is often a reliable signal that the blamer takes the norm seriously.

A major strength of this account is that by explaining how blame can signal different things, it successfully accounts for why blame takes so many different forms. We can signal commitments and willingness to enforce norms by experiencing strong emotions or reactive attitudes, modifying our relationships with norm violators in response to those violations, protesting a norm violation, or communicating to norm violators (either verbally or non-verbally) that we find their behaviour unacceptable. Such a signal is costly because blaming stirs up unpleasant emotions, and motivates us to act in ways that involve risk. Blaming you risks retaliation, or that a tension in our relationship will result, or that our relationship will dissolve altogether. Another potential cost consists in the costs one would incur if their blaming were to be deemed inappropriate by the rest of the moral community (Fraser, 2012). This fact can contribute to others' understanding of what someone is signalling. If it is common knowledge that people who blame too strongly are themselves met with a withdrawal of good will or sanctions from others, then when I blame strongly, I communicate that I take this norm violation to not be a trivial matter. Alternatively, I imply that I take this matter seriously enough that even if others would disagree with my assessments, I am willing to bear the costs of those disagreements.

Though I take the insight that blame is a costly signal to be a significant contribution to our understanding of blame, I think there are three shortcomings with this account. The first is that their proposed explanation for blame's contribution to norm-following via signalling is incomplete. Shoemaker and Vargas (2021) emphasise how both observers and signallers gain from this system via it solving "the many everyday prisoner's-dilemma-type situations we find ourselves in" (587), primarily from the fact that being a reliable signaller and upholder of a norm are good grounds for a trustworthy reputation. If I know you have a reputation for not cheating, I can count on you to not cheat in situations where relying on you gives you the power to exploit me. Though this is correct, they leave out how the benefits the signalling system generates doesn't just make blame strategically rational for the signaller and

helpful to the observer in the long run; they also valuably contribute to norm-maintenance within the wider community in both the short- and long term. Seeing how requires moving from thinking about prisoner's dilemmas to mixed-motive public goods games and understanding how assurance can be provided independent of reputation. Recall that in these settings, players behave like conditional co-operators: they want to secure the highest collective payoff and so want to invest their resources, but they'll typically only do so if they believe others will also contribute. This is how blame *qua* costly signal of commitment enables cooperation. In blaming, I signal that I am committed to this norm, giving *assurance* to others they will not be taken advantage of if they co-operate, and to potential defectors that norm violations will not be tolerated (elaborated on below). My blame signals what the norms are like 'around here', or as seen by people like me, even if you'll never encounter *me* again (cf. norm-enforcement in online spaces). Additionally, expressed blame helps make both the expectation that norms will be followed, and the fact that particular transgressions have occurred, common knowledge (cf. Pinker, 2025). Even reputation-based effects thus aren't limited to that of the blamer: expressing blame also often diminishes the target's reputation, helping other norm-followers to avoid situations where they could be taken advantage of by the violator.[14]

The second shortcoming is that when we consider blame within our moral accountability practices, this explanation seems to miss the phenomenology and our self-understanding of what we are doing. McKenna (2024) asks us to "imagine that someone was clear-eyed in a moment of directed blame in such a way that she might announce as she blamed another "what I am most fundamentally doing is signaling that I am a member of a particular moral tribe" (292). This would not seem to be an adequate description of what is occurring, or of what most people take themselves to be doing. It also seems to generate normative worries about appeals to the wrong kinds of reasons, because such responses omit any reference to the actual conduct of the norm-violator.

This relates to the third worry, which is that it that if blame's exclusive function is signalling, we fail to answer why blame has a sting, opprobrium, or why inapt blame could ever be unfair or inappropriate. As noted earlier, inappropriate blame seems to involve more than just a lack of fittingness, understood here as correctly representing its object (in blame's case, as a culpable wrongdoer). Blame is somewhat special in this regard; if directed towards someone who isn't blameworthy, it isn't just unfitting, it is a form of *mistreatment*. But if blame's function is simply to signal internalisation, endorsement, and willingness to enforce (in contrast to actual enforcement), it is puzzling why we would ever object to blame. After all, one can credibly signal such qualities in a variety of ways, such as writing lengthy essays and raising awareness, but they don't seem like blame, or like the kind of thing that could generate the kind of objections we level at inapt blame, precisely because they lack blame's distinctive sting. As a result, while this theory does fantastically at accounting for the disparate variety of phenomena that all count as instances of blame, it doesn't yet account for

[14] This focus on reputation is also present in Shoemaker's (2024) later treatment of the theory, which again emphasises how blaming is strategically rational, somewhat sidelining how the benefits to observers also contribute to norm-following.

the main property that made us interested in blame in the first place. To understand blame's sting, we need to introduce a second effect.

4 Sanctioning

We can begin to understand blame's force by noting something that Shoemaker and Vargas's account misses: in many cases, blaming *just is* a way we don't tolerate someone's norm violation.[15] Being blamed is often unpleasant. It can induce guilt in its target, understood as a pained recognition of what one has done (Carlsson, 2017). If blame takes the form of relationship modification, as Scanlon (2008) takes it to, one loses out on many things that were previously valuable. When we express our blame to third parties with the transgressor absent, we diminish their reputation. Often we don't just signal what we will do if the violator doesn't change their behaviour, we *threaten* what we will do (Reis-Dennis, 2019). And in publicly blaming someone, we can change the payoffs of various behaviours they were considering. For example, now that everyone knows what you've done, and has seen how seriously that transgression affected me, failing to apologise risks that others will stop interacting with you because they don't want to signal they're not on my side. These observations are not only relevant to identifying phenomena that our theory of blame ought to account for, they hint at a second important benefit that blame delivers.

On Shoemaker and Vargas's thinking, agents are first committed to norms and that commitment disposes them to enforce said norms. In response to norm violations, our commitment sometimes causes us to experience angry emotions which then help motivate us to enforce them (587). This account thinks of such emotions as part of our blame, but takes the blaming (the emotions) to be conceptually distinct to the enforcing.[16] While this separation can be theoretically useful given one's interests, reserving the term 'blame' exclusively for the former doesn't quite map onto how we ordinarily describe many interactions featuring blame, and more importantly, doesn't capture the way in which we typically object to someone's blame. Again, in objecting that someone's blame is too harsh, we are not objecting that the agent's signalling of their commitment is too accurate or too reliable, or acting as too strong a piece of evidence that they are committed. We are also not just objecting that this response is unfitting, the way that we might object to other unfitting reactions. We are clearly objecting to the agent's *treatment* of their target, i.e., the way that they are enforcing the norm or not tolerating the target's behaviour.

We can find an additional basis for tightening the link between blaming emotions/attitudes and enforcement behaviour by noting that enforcing behaviours produces similar benefits to signalling behaviours, which were used to help motivate the idea that signalling is a part of blame's function. Let's return to the public goods game

[15] Shoemaker's (2024) later treatment is better on this point; the details are examined below.

[16] They allow that accounts of altruistic punishment are complementary with their own (p. 588), but think that explaining the diversity of blame phenomena requires giving the signalling aspect priority, even though punishing behaviours help explain the signal's reliability. I am claiming that explaining the sting requires giving priority to what helps explain the signal's reliability, which is the disposition to actually sanction, and that this cannot be separated from our understanding of what blame is.

with conditional co-operators. Signalling alone doesn't always secure continued cooperation, because some people do not want to secure the highest collective benefit even if everyone else is doing so. These defectors pose a risk in two ways. First, by taking and not contributing, they reduce the total payoffs available for everyone. But secondly, if co-operators see defectors taking advantage of the contributions that everyone is helping produce, they become much more reluctant to continue co-operating. This is because conditional co-operators do not always switch from cooperating to defecting at the same point as each other. If I stop cooperating when I see three people defect, this will then trigger anyone who would stop cooperating upon seeing four or more people defect to stop cooperating, which will then trigger anyone who would stop upon seeing five or more … etc.

This is where sanctions come in. Sanctioning increases the costs of—and thereby deters—defecting.[17] If it is applied early to small numbers of defectors, it can prevent a cascade of defections from developing (Fehr & Gachter, 2000).[18] It particularly helps to reduce incidences of defection if potential defectors know ahead of time they will be sanctioned, rather than if they only learn this after they try defecting, which is another way that signalling provides assurance. But in order for would-be defectors to know this—in order for agents to be able to signal that norm violations will not be tolerated—the signals *must* have a sufficiently reliable connection with *actual* dispositions to enforce. We cannot understand blame's signalling function *qua* willingness to follow and enforce the norm without presupposing that blaming agents (as a group) are *in fact* disposed to exhibit norm-enforcing behaviour.[19]

[17] Shoemaker (2024) proposes that blame's sting provides a 'jolt', which serves "a reminder of how we deviated from the norms and an incentive not to do it again" (28). However, he provides little explanation of *why* it has this jolt, and doesn't outline the importance of preventing cascades. (He later mentions that it typically produces certain negative emotions, such as shame or guilt (128), but it seems blame's sting outstrips simply generating emotional reactions in the target, as other means of producing these emotions don't have the same sting). Importantly, he also takes this 'jolting' to be conceptually and normatively independent of the 'sanctions' he discusses later, which also don't depend on blameworthiness to be fair to dish out. He does grant that blaming emotions can be expressed with sanctions, but emphasises that most blame involves no sanctions (156), making it unclear what causes e.g. stronger, angry blame to have more of a 'jolt' than dispassionate blame even if they produce the same emotional response in their target and manifested in the same kinds (or absence) of sanctions. My proposal provides an explanation for why blame jolts, why certain types of blame characteristically jolt more than others, why treatment is at issue even if blame is private or has no effect on its target and even if emotions and attitudes seem central, and why it seems like blaming is primarily about responding to transgressions, rather than signalling facts about the blamer.

[18] Strictly speaking, this is usually described as 'punishment', but 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. 'Sanctions' is a more useful category for us for a few reasons: first, it overlaps with how such behaviours are already described elsewhere in the blame literature. Second, I do not intend to commit to the kinds of classifications that biologists are interested in; whether we wish to consider a behaviour sanctioning doesn't seem to depend on whether the behaviour is immediately costly or fitness enhancing, for instance. Third, 'punishment' has a whole range of additional associations in philosophy (e.g. with institutions, that the infliction of suffering be intentional) which we do not need to draw upon.

[19] Shoemaker's focus on blame's propensity to 'jolt' (independent of 'sanctions') leads him to sometimes conflate the signal and the hard-to-detect property. For example, he says: "To the extent that blamers signal these commitments, those who pick up on the signal need a way to know that it's reliable, that it's hard to

This close link between blaming attitudes and sanctioning behaviours used to receive much more emphasis.[20] Strawson (1962) sets up the debate between the optimists and pessimists as concerning our "practices of punishing and blaming, of expressing moral condemnation and approval" (p. 1), mentioning 'punishment' separate to condemnation no less than nine times. He also makes clear towards the end of *Freedom and Resentment* that some punishment-like responses remain within the scope of his account, being "all of a piece with this whole range of attitudes of which I have been speaking"

fake, that it's honest" (32). But if blame is the costly signal, blaming *just is* the reliable thing that lets observers infer signallers have the qualities they are interested in. Similarly, he says "The information signaled by norm-maintenance-via-blame is incredibly valuable (and otherwise very difficult) for others to garner, as it reveals the blamer's commitment to enforcing norms" (32). The quality of interest is thus whether blamers will "deliver blame's sting to [norm-violators] as well. It's that information that contributes most directly to norm maintenance" (33). But propensity to blame or 'enforce norms' (via jolting) is very observable when someone is blaming; an assumption that past behaviour predicts future behaviour will suffice for observers. What is much harder to observe—and can be signalled via blaming—is a disposition to reliably follow the norm (especially in rarer, higher-stakes circumstances), and caring enough about transgressions to be willing to police the norm via *increasingly harsh* sanctions. Shoemaker also claims that blame which seems costless to signal (e.g. private blame) remains costly because there are 'competence costs' involved: those required to become an honest signaller (34) who has learned, internalised, and knows how to enforce such norms appropriately. While it is true such costs are involved in becoming an *honest* signaller, these aren't what *make* blame a 'costly signal' because if dishonest blaming were cheap, blame would still be a useless signal for observers. What he means to describe are the costs inherent to acquiring *the hard-to-observe quality* of interest, which I take to be commitment, rather than propensity to enforce *qua* jolting with blame. All of the work is being done here by the fact that beings without the quality of interest would be unlikely to signal simpliciter (not that there were costs involved for those who do signal). Recognising this important detail helps avoid several other objections that his explanation is open to. Blame isn't a costly signal in virtue of the fact that honest blamers have incurred a lot of costs (young children are very capable of blaming without having incurred such costs, for instance, despite also not being reliably committed to many norms), or that they're competent at blaming (competence at blaming is the kind of thing that can be assessed when observing me express blame, and it seems unlikely that observers need some way to know who was competent at blaming in the first place) or that they're competent with 'the norm system' more generally (we don't see someone blame thieves appropriately and conclude e.g. they also know how to blame adulterers appropriately). Rather, it is in virtue of the fact that people who do incur such costs tend to care about certain norms in ways that people who haven't incurred such costs don't, which affects the chances they will blame simpliciter.

Shoemaker also claims that blame which seems costless to signal (e.g. private blame) remains costly because there are 'competence costs' involved: those required to become an honest signaller (34) who has learned, internalised, and knows how to enforce such norms appropriately. While it is true such costs are involved in becoming an *honest* signaller, these aren't what *make* blame a 'costly signal' because if dishonest blaming were cheap, blame would still be a useless signal for observers. What he means to describe are the costs inherent to acquiring *the hard-to-observe quality* of interest, which I take to be commitment, *not* propensity to enforce *qua* jolting with blame. All of the work is being done here by the fact that beings without the quality of interest would be unlikely to signal simpliciter (not that there were costs involved for those who do signal). Recognising this important detail helps avoid several other objections that his explanation is open to. Blame isn't a costly signal in virtue of the fact that honest blamers have incurred a lot of costs (young children are very capable of blaming without having incurred such costs, for instance, despite also not being reliably committed to many norms), or that they're competent at blaming (competence at blaming is the kind of thing that can be assessed when observing me express blame, and it seems unlikely that observers need some way to know who was competent at blaming in the first place) or that they're competent with 'the norm system' more generally (we don't see someone blame thieves appropriately and conclude e.g. they also know how to blame adulterers appropriately). Rather, it is in virtue of the fact that people who do incur such costs tend to care about certain norms in ways that people who haven't incurred such costs don't, which affects the chances they will blame simpliciter.

[20] Cf. Mill's (2015), p. 165 famous remarks: "We do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, by the opinion of his fellow creatures".

(p. 23).[21] In particular, "the other-reactive attitudes are associated with a readiness to acquiesce in the infliction of suffering on an offender" (p. 24). While 'inflicting suffering' might be too strong for most readers, it certainly seems like such attitudes provide a readiness to engage in negative treatment, i.e. to sanction.

Similarly, Watson (1996) believed that holding accountable "involves the idea of liability to sanctions" (p. 237).[22] He proposed that blaming attitudes are disagreeable because they involve "dispositions to treat others in other generally unwelcome ways" (p. 238), and this makes the attitude itself subject to fairness considerations, requiring some justification for infliction: "If it *would be* unfair to impose certain penalties on someone, then it is unfair to be ready (however impotently) to do so" (p. 239).[23]

Finally, Wallace (1994) also saw a close connection between blaming attitudes and sanctioning behaviour in his influential treatment.[24] He considered holding people responsible to be a stance that agents take, "characterized by the responses of blame and moral sanction" (p. 11). To blame is to be subject to one of the reactive attitudes, and these attitudes express themselves in sanctioning responses, "such as avoidance, reproach, scolding, denunciation, remonstration, and (at the limit) punishment" (p. 53). While, strictly speaking, he took blame to be an attitude, to emphasise this while omitting his continual pairing of blame with sanction as a single object (fifty-four times, by my count) seems like insisting on the letter of his account while ignoring much of the spirit.

Why is it that so many people who build on Strawson, Watson and Wallace give such disproportionately low attention to sanctioning behaviours compared to attitudes and emotions?[25] According to Smith (2013, p. 30), "While this sanction account of

[21] "The concepts we are concerned with are those of responsibility and guilt, qualified as 'moral', on the one hand—together with that of membership of a moral community; of demand, indignation, disapprobation and condemnation, qualified as 'moral', on the other hand—*together with that of punishment*. Indignation, disapprobation, like resentment, tend to inhibit or at least to limit our goodwill towards the object of these attitudes, tend to promote an at least partial and temporary withdrawal of goodwill; they do so in proportion as they are strong; and their strength is in general proportioned to what is felt to be the magnitude of the injury and to the degree to which the agent's will is identified with, or indifferent to, it. (*These, of course, are not contingent connections*.)" Strawson (1962, pp. 23, emphasis added)

[22] Though he earlier (Watson, 1987, p. 148) expressed some unease with the retributive elements of Strawson's account, arguing that it was possible (though rare and difficult) to hold accountable without reactive attitudes.

[23] Watson is less than definite on whether the attitudes themselves are sanctions. On the one hand, his discussion concerns the fairness of sanctions, and he agrees that it can be unfair to blame the dead, implying that the blame would therefore be a sanction. On the other hand, he says that blame cannot be disagreeable for the dead as a sanction, because they cannot be subject to adverse treatment. This seems to suggest he thinks that blaming attitudes can be unfair, but are not themselves sanctions. But one might question this: if the fact that x would be unfair is enough to make the readiness to x unfair, why not think that the fact that x is a sanction is enough to make the readiness to x a sanction?

[24] While Wallace (1994, p. 54–59) argues against an 'economy of threats' view, Hieronymi (2004, fn. 24) notes that he doesn't seem to get very far from it. The importance of sanctions is oft-emphasised in passages such as: "Sanctioning behavior belongs to the syndrome of responses to which the reactive emotions dispose those who are subject to them, because the connection with reactive emotions is part of the conventional meaning of such behavior. We learn the concepts of indignation, resentment, and guilt in part by learning to see their connection to sanctioning behavior, and the adequate expression of those emotions often requires such behaviour … blame and moral sanction can be seen to have a positive, perhaps irreplaceable contribution to make to the constitution and maintenance of moral communities" (p. 67–68).

[25] Some representative examples: Sher (2005): "if we wish to understand blame, it is not these expressions, but rather the attitude they express, upon which we must focus" (p. 74). Graham (2014, p. 391):

blame was once widely accepted in the philosophical literature, I think it is fair to say that it has now fallen decisively out of favour," partly in backlash to utilitarian approaches to blame (Schlick, 1939; Smart, 1961). She notes two objections to thinking of blame as a kind of sanction: first, it is hard to see how private blame can sanction, given the target never finds out about it. Second, it is hard to see how we can sanction the dead or distant, and yet we clearly can blame them.

These points aren't strictly false, but their relevance is disputable.[26] I grant that private blame doesn't seem like a sanction if we just assess its 'sanctionness' according to whether or not every single blamed target's interests are, in fact, set back. But this is the wrong way to think about blame's nature; it is analogous to concluding that prison isn't best understood as a form of punishment simply because there are some agents whose interests would not be set back by being put in prison. Even if I cannot, in fact, sanction the dead, my blame makes me prepared to: I am clearly *more disposed* to do so than I would be if I were not blaming them at all (McKenna, 2016). Smith and many others seem to think that since we can experience blaming attitudes without expressing them or sanctioning, these attitudes have explanatory priority or are somehow more basic than the behaviours. My argument is this is a mistake: our preparedness to sanction and our attitudes are part of the same underlying reaction. *Such reactions only function as reliable signals of our unwillingness to tolerate violations because of their reliable connection to these sanctioning behaviours*.[27] To be clear, sanctioning dispositions can be masked by other factors: a worker might not be very disposed to sanction their boss despite blaming them, due to fear of being fired.

"As blaming someone consists in feeling certain emotions toward her, it is not a form of adverse treatment at all". Talbert (2012, p. 90): "When I state that a person is blameworthy, I mean that attitudes like resentment are reasonably directed at that person, regardless of whether anyone experiences or expresses these attitudes … in some cases the mere judgment that a person is a proper target for these reactions also constitutes a form of blame." Shoemaker (2024, p. 156): "Most of our blame/praise system actually doesn't involve sanctions." Several philosophers get more into enforcement territory when they speak of blame's demands, but they tend to see these as coming from attitudes or emotions, rather than sanctioning, and such views are not without problems (see Macnamara, 2013). Some philosophers come closer to emphasizing sanctions, but still take them to stem from emotions or attitudes. Carlsson (2017) argues that the blameworthy deserve to feel guilt, understood as a pained recognition that one has acted with ill will, and that insofar as resentment provokes this, it acts as a sanction, with similar sentiments from Portmore (2022). Rosen (2003) thinks blame is a sanction, but only takes it to be an emotional response: "Even when it is not expressed, it is a form of adverse treatment: a form of psychic punishment" (p. 73). King (2014, p. 412) proposes that while holding accountable does involve sanctions, claims about the desert of blame concern attributability responses. McKenna (2012) is an exception: "What is more fundamental to the nature of [moral] emotions … is the public manifestations of them. Our understanding of the private cases is parasitic on an understanding of the paradigmatic public ones" (p. 69). Sceptics about basic desert (e.g., Pereboom, 2021; Caruso, 2021; Menges, 2021), do focus on forms of blame that are more sanction-like, though given they are happy to revise our practices as we find them, they aren't seeking to build on Strawson, Watson and Wallace.

[26] She also objects that though we sanction children we do not blame them, which seems to be attributable to the fact that our actions do not express the same attitudes. This is more contestable: even if blame were a sanction, this doesn't entail that all sanctions are blame, and many people *do* blame children, they just find it inappropriate upon reflection.

[27] This gives a stronger foundation to McKenna's (2012) contention that private blame 'is parasitic' on its public form, which had been questioned (Russell, 2017). Our account also has a much easier time explaining why children can blame before they understand moral conversations, which acts as an objection to McKenna's conversational model of blame (Driver, 2016).

But there are always some ways that blaming agents are more disposed to treat their targets negatively, or at least, are less prepared to treat them with good will.[28]

We can put pressure on the relevance of examples supporting a divorce between blaming attitudes (and intentions and expectations, in Scanlon (2008) and Smith's (2013)) case and blaming behaviour by trying to imagine someone who experienced blaming attitudes (intentions, expectations, etc.) such as indignation, but who was not disposed *at all* to change their behaviour towards the target in any way approximating sanctions, ever. Such a being would be incredibly psychologically abnormal, and I struggle to imagine them as truly experiencing indignation, as opposed to being confused about experiencing some other negative feeling.

In the other direction, consider two agents who equally believe we are blameworthy for a norm violation, and who treat us negatively in identical ways in virtue of that fact. Suppose we are aware that the only difference between them is that one experiences blaming attitudes such as resentment. The other does not, but they are nevertheless committed to the norm we violated, they care deeply about it, and enforce it accordingly. Is it really the case that the latter's actions entirely lack the 'depth, force or sting' that we commonly take to be at issue in blame?[29] If these agents had good evidence that we were wholly innocent but did not change their behaviour or beliefs, would our objections to each agent significantly differ due to the fact that only one had a certain attitude?[30] It also doesn't seem like we would make two distinct objections to the agent who had the attitude and the negative treatment, but only one to the agent who lacked that attitude. We would object to both of them on the basis that they are unfairly enforcing a norm that we did not culpably violate.

Such examples, I think, show it is a mistake to infer that since we can have blame-related attitudes or emotions without sanctioning behaviour, we can understand blame (and with it, practices of holding accountable) while omitting reference to sanctions entirely.[31] If one wants to insist that 'blame' should be reserved for the attitude, then I would propose we consider norm-enforcing expressions of this atti-

[28] Cf. Wolf's (2011, p. 336) dispute with Scanlon (2013, p. 99) about whether she can blame her daughter without their relationship being impaired at all. On my account, we can sidestep debates about how to best understand intentions and expectations in relationships, and note that while blaming she *is* disposed to treat her daughter more negatively in some ways, even if this does not affect 'the relationship' overall.

[29] McKenna (2012, p. 113) shares my scepticism.

[30] To be clear, I am not claiming there would be no difference. Just that this difference is likely to be relatively small, and this puts pressure on the claim that the attitudes alone are doing the work we attribute to blame, with the actions simply being a downstream vehicle of their expression.

[31] A similar divorce occurs in the norms literature, and this points to a second reason why the norm literature and blame literature have taken such little notice of one another. Bicchieri and Brennan et al. (2013) both take pains to outline how some norms (e.g. eating with one's fork in their left hand) can be maintained without sanctions, and how people do not decide to follow norms based on whether there are sanctions, often acting automatically. However, this is simply a reflection of different interests. Bicchieri, for instance, is largely interested in accounting for the *relative variation* that occurs in rates of norm-following when sanctions are held constant. But I am not claiming that *no* norms can be upheld without sanctions, or that sanctions are the *only* way to uphold norms. My interest is in the *stability* of norms, particularly moral norms, which are the most difficult to sustain without sanctions. Brennan et al. (2013) agree that norms favouring cooperation need some kind of sanctioning mechanism to be upheld (p. 137), and Bicchieri later notes, "social norms are closely tied to sanctions; without sanctions (internal or external), they may not exist" (Bicchieri, 2014, p. 212). Even if some norms don't require sanctions, these accounts do not show

tude to be instances of 'blam*ing*', and take 'blaming' to be the more important part of our moral responsibility practices and how we hold agents accountable.[32] Emotions and attitudes are something we experience, but moral practices are something we participate in; blam*ing* and hold*ing* accountable is something we *do*. This would help highlight the mistake made in taking there to be a merely contingent—correlated but theoretically inessential—relationship between blame and blaming such that one could understand the former without careful consideration of the latter.

Having tidied up the way that we think about blame's sanctioning qualities, shown that taking blame to be a sanction significantly improves Shoemaker and Vargas's signalling-focused account, and rehabilitated a line of thought with a considerable historical pedigree which used to be much more recognised, we have made considerable progress towards illuminating blame's nature. However, the complete picture requires understanding one last effect.

5 Sensitising

Suppose we have a community of conditionally-cooperating agents who are assured both that others will cooperate and that defectors will be punished. This may be sufficient to maintain valuable forms of norm-following in a highly idealised world. But in the real world, uncertainty is a fact of life. We are often uncertain that everyone has our backs, or that defections will be punished, as are potential defectors, who may take their chances promoting their selfish interests. Additionally, the conditions encouraging others to comply with certain norms can be precarious, and sometimes agents accidentally send the wrong signal. There can also be uncertainty about what kinds of behaviours count as complying with the norm, and the circumstances in which the norm does or does not apply. Things are particularly difficult when we have multiple norms to be complied with, which can conflict with one another, and live within multiple communities with different sets of norms.

What would help compliance is if we had groups of agents who were motivated to comply with shared norms without needing much assurance, or who would do so even if the chances of punishment for violating appeared relatively low. If such agents could find each other (including by gaining evidence about each other's reputation; Sperber & Baumard, 2012), they could cooperate without continually having to evaluate whether others were likely to think they can do better by defecting. If such agents were numerous, they would also be able to share the costliness of monitoring for violations. They might not even have to actively monitor at all, they could just go about their day and sanction whenever they happened to encounter a violation. In short, what would be useful is if we had agents who had *internalised* the norm: agents who endorsed the norm, and felt *inherently* motivated to both comply with and enforce the norm. This, I propose, is the last important effect blame has, being another means by which it greatly

they are likely to be particularly stable or to have been developed without sanctions (e.g. regarding table manners generally).

[32] Vargas (2013, p. 119) also proposes a distinction between 'blame' and 'blaming', but the former refers to a minimal judgement of blameworthiness while the latter is an emotion-laden reaction, which may just be having a reactive attitude.

enhances cooperation. Though blame from others (*qua* sanction) starts out for agents as a source of external motivation to comply with norms, as agents continually try to comply (particularly as children) they come to internalise the norm.[33]

Internalisation is particularly useful because it produces agents whose motivations to comply with a norm outlast their external, contingent incentives to do so: it makes their cooperation *robust*. A community of agents who have internalised the same norms don't have to pay the costs of continually calculating whether it's in their interests to cooperate or defect, or whether other agents are likely to engage in similar calculations. Just as worries about defection can become a self-fulfilling prophecy, making defection in fact more likely (cf. inflation; bank runs), confidence in others' cooperation can also be self-fulfilling. Agents who have internalised norms don't need to waste resources on 'insurance' to cover instances where they get taken advantage of, or 'security' to prevent being taken advantage of in the first place. They can significantly reduce the need to gather information about others and monitor for violations, and they can minimise the costs spent on developing enforcement mechanisms.

Interestingly, agents who sanction due to experiencing norms as inherently worth upholding can actually be *more* effective at deterring norm violations than agents who explicitly sanction to deter. One of the known problems with deterrence-based consequentialist justifications of punishment is that they fail to justify punishing agents who would commit one-off violations. In game theory, strategies that sanction norm violators in order to deter future violations also face a similar problem. Perhaps *this* agents' violation was a one-time thing, or perhaps now that the agent has committed the violation and learned what the experience was like, they have lost all motivation to ever try it again. If the point of the sanction is deterrence, and this—by stipulation—is an instance in which there is no need to deter the agent, then we have no reason to punish the violator in this instance. Even if publicly sanctioning here would arguably still deter others, this could theoretically be achieved by simply pretending to sanction, or having the violator agree to not tell anyone that they were let off the hook. But if would-be violators are aware of this possibility, they will not be

[33] Praise for meeting certain standards also encourages internalisation of certain behaviours. The insight that blaming and praising practices can produce internalisation has been proposed by Vargas (2013, p. 175–177), and my treatment here is heavily indebted to his account, in addition to his and Shoemaker's (2024) paper. But his focus is on reasons-responsive agency, rather than norm-following, and despite their overlap (since reasons-responsive agents are necessarily disposed to comply with moral reasons which would manifest as moral norm-following) there are distinct advantages to focusing on the latter. First, as already shown, it more explicitly connects up with the literatures on norms and cooperation. This is notable because if blame's function was to produce *moral* reasons-responsiveness, we would have to say that blame which produces agents who comply with a suite of unjust norms has thereby failed to fulfil its function. But an account of blame's nature which explains internalisation of *both* good and bad norms (the latter of which is orthogonal to moral reasons-responsive agency) has more explanatory power. Second, Vargas locates the value of blame in its effects on the blamer. But we don't blame people for *their* benefit; we blame them for *our* benefit (to appropriate Hieronymi (2019): "I'll bet you think this blame is about you [but it isn't]"). Locating the benefits of blame in norm-following is much more faithful to this. Third, since this story was developed prior to his work with Shoemaker, he needs some kind of bridging story to avoid having two distinct accounts: one about blame's nature (a signal of commitments) and one about its justification (being unpleasant and thus a motivator to produce valuable agency). My account is more parsimonious on this front: given our norm-psychology (discussed below), blame's signalling and sanctioning effects just are what produces beings who comply with norms, and this makes our stories of blame's nature and justification continuous with one another.

deterred against violating the norm in instances where they know that post-violation sanctions would not have an effect (or worse, when they merely believe that they can convince others that deterring is not needed).

One way community members can prevent these kinds of violations from occurring is through possessing dispositions which act like forms of *pre-commitment* (cf. Frank, 1988).[34] If it is known that violators will be punished *regardless* of whether the subsequent sanctioning seems to deter them, community members can thereby actually create more deterrence.[35] Agents who have internalised a norm, such that they care about it for its own sake (rather than simply as an instrument for ensuring co-operation), act as a credible threat that violations will be sanctioned even when deterrence seems unneeded. Put simply, potential defectors won't try to find ways of violating the norm without being sanctioned if they know these exceptions are not available.

Our tendency to internalise and enforce norms is the result of our 'norm-psychology'. A substantial body of evidence shows we possess "a suite of genetically evolved cognitive mechanisms for rapidly perceiving local norms and internalizing them" (Chudek et al., 2013, p. 443; see also Chudek & Henrich, 2011; Davis et al., 2018; Sripada & Stich, 2006). This is one of the features that makes us distinct from other animals, including chimps (Tomasello, 2015). Norms are a cultural universal, and "in all human groups, systems of sanctions, which utilize ostracism and gossip, as well as other informal sanctions, are applied to those who violate moral norms" (Sripada & Stich, 2006, p. 287). Despite the content of norms displaying high cross-cultural heterogeneity, most of this variation in adults is already present by the time children are nine years old (Henrich et al., 2001). Young children, for instance, spontaneously infer the existence of social rules, react negatively to transgressions by others, and sanction transgressors, even in single trial learning (Rakoczy et al., 2008; Schmidt et al., 2016). That they spontaneously and punitively enforce norms on other children even without having seen others do so (Edwards, 1987) suggests the mere observation of norms being followed induces certain normative inferences. Norm-following is often automatic, requiring overriding of more habitual responses, and children even perform better on deontic rule reasoning tasks than they do on otherwise similar reasoning tasks. Both adults and humans preferentially attend to and learn from individuals with greater skills, confidence, experience, prestige, and ethnic markers matching their own, things which are likely to enable one to more quickly learn 'how things are done around here' (Chudek & Henrich, 2011).

We now have the resources to address the earlier objection that Shoemaker and Vargas's signalling-based explanation seems to underdescribe what agents are *doing* when they blame. Blame is indeed a costly signal insofar as it's a signal that would be unlikely to be emitted by agents who are not committed to norms, as they wouldn't care enough about the norm-violation to be willing to take on the costs involved in blaming. But this is not *all* that blaming agents are doing; particularly when expressing their blame to the norm-transgressor, they also *just are* enforcing the norm, which

[34] From *Games of Strategy*: "Tying yourself to a rule you would not want to follow if you were completely free to act at a later time is an essential part of the process" (Dixit & Skeath, 2015, p. 329).

[35] Note that one can pre-commit to blaming *culpable* norm violators; we are not forced to blame violators who turn out to be young children, for instance.

is where a significant proportion of those aforementioned costs obtain. Adding the norm-psychology into our story then accounts for the phenomenology and can generate the right kind of reasons: because we feel inherently compelled to uphold the norm, we feel inherent motivation to sanction those who violate it: that is, our motivating reason for blaming will often be *that they seem blameworthy*.

Introducing norm-psychologies into our story accounts for another important way that blaming contributes to norm-following, which can be easy to miss. This body of research demonstrates that humans are not just primed to develop motivations to comply with norms, we are also particularly adept at *learning* what the norms are and how to comply, and it is useful to keep these threads distinct.[36] When others blame, a number of effects occur concurrently with my increased motivation to comply with the norm. I am receiving *information*, not just about you, but also about what the norms are like around here. In being blamed myself, I am also receiving *feedback* on my behaviour. And as I continually try to comply with the norm—even if only at first to avoid external sanctions—I am *practicing*. Over time, the conjunction of these things develops my skills, making me become *competent* with the norm.

When considering the stability of norm-following, the importance of developing competent agents can be easy to miss for three reasons. First, game theoretic models regularly take it as an assumption that players know what their options are and how to choose them (typically: cooperate or defect). Second, as already noted, competence often develops alongside the motivation resulting from signalling and sanctioning. Third, many paradigmatic moral norms seem very obvious to us and easy to comply with (e.g. 'don't murder'). But such reasons are clearly insufficient to warrant thinking that accounts of norm-following do not need some understanding of how norm-competence develops. Game theoretic models are explicit idealisations, and we clearly consult our intuitions about the obviousness of various norms and means of complying *after* we've had decades' worth of training with said norms. A realistic account of norm-development must recognise that agents need to comply with norms that are highly variable across cultures and contexts, which involve multiple sets of norms and interests that can conflict with one another, and which do not wear on their sleeve which actions count as violations. Norms require quite a lot of training and know-how to navigate, and our norm-psychology makes us particularly sensitive to norm-relevant information.

That competence is a distinct and valuable property to produce can be seen by observing that deterring violations through sanctioning does not always ensure valuable forms of norm-compliance.[37] For example, imagine an agent who is faced with an unfamiliar domain that has very subtle norms, e.g., a particular arena of politics, or a field of employment, or a particular community. Suppose that one could learn how to navigate this domain with time, education and practice, but that violations are very costly. Many agents who lack competence with these norms will plausibly, due to fear of being sanctioned, simply choose to avoid that domain altogether. Here

[36] This is a detail that is somewhat missing from Vargas's treatment; though he is interested in reasons-sensitivity, and is sensitive to the ways that our environments can sustain or enhance different forms of agency, he takes continued attempts to avoid blame to primarily affect our motivation to comply with norms.

[37] Such effects can be modelled by modifying public-goods games from the standard treatment presented earlier.

the threatened sanction has produced the effect of preventing norm violations, but only at the cost of also preventing a certain kind of norm-uptake. The agent will not have gained competence with the relevant norms: they do not sufficiently understand them, cannot explain them, follow them, weigh them, or recognise them. Such outcomes will be especially bad if we want people to ultimately become familiar with the norms, perhaps because they solve some important problem.

There is a difference between sanctions simpliciter, aimed at deterring behaviour, and negative responses which are targeted, proportionate, meaningful, allow an opportunity for reply, and made with an eye towards the agent's long-term development. Our blaming responses frequently take such forms: it is widely-accepted that blaming should not be disproportionate to the wrong, that our reasons to blame are affected by how the wrongdoer responds to their wrongdoing (e.g. decreasing when they apologise and signal commitment to improvement), and that there is little point blaming when the agent cannot understand what they did wrong. Such constraints on blaming help develop the target's competence. Particularly as children, or when operating in a domain that one has not had much familiarity with (e.g., visiting a new culture), we try out different actions and learn what kinds of actions count as violating the norm and which ones don't. We also learn about norms that apply only in particular contexts or to particular agents: those that occupy certain roles, for instance. And we learn how norms compare to one another and how to handle instances in which they conflict, such as whether some third alternative is required, or whether one is to be weighted more heavily in our deliberations. At times we will overgeneralise, while at other times in which we have been cautious we will notice that other people's actions were not penalised, showing we can loosen up a bit.

Once this norm-psychology (and a suite of other capacities) are sufficiently developed, blaming responses can sensitise agents to moral considerations via several routes. In blaming you, my blaming behaviour immediately draws your attention to your conduct, and in particular, your conduct as seen by others. Note that this blame doesn't have to be directed at you: simply finding out from others how much I blame you behind your back can be enough to prompt reflection, and an awareness that I can't be assumed to treat you with good will. Then, if we begin dialogue, we enter into an (often heated) conversation about what you did and the significance of that behaviour, further prompting you to notice considerations you may not have noticed before. In addition to the reasons I am giving you, my emotional responses can produce emotional responses in you such as guilt or shame, via the engagement of your empathy. As I blame, you see how I experienced your actions. Although you don't feel the exact same response as me via empathetic contagion (whereas I feel victimised, disrespected, or insulted, you feel guilt and regret), your responses are generated in response to an empathetic recognition of what I am feeling and how I experienced the wrong.

The relevance of blaming practices for developing the norm competence of agents is greatly emphasised by McGeer (2013, 2015, 2019). She argues our moral agency is a *skilled* competence, which can be 'reactively scaffolded' by the moral community, particularly when there is a back and forth between wrongdoers and others to develop moral understanding. As she puts it: praise, blame, punishment and reward "are essential if we are to develop and sustain the very capacities on which successful norm-governed thought and action depends" (2015, p. 273). This story also has some sympathies from Fricker (2016), who focuses on blame's ability to act as a prolep-

tic mechanism. By treating the violator as if they can recognise the moral reasons they have to e.g. apologise and change their ways, even though they don't presently, we thereby produce that very capacity and recognition, bringing their "moral understanding into alignment with our own" (p. 178).

Though McGeer (2013) and Fricker (2016) give some acknowledgement of how reactively scaffolding or aligned understanding involves motivation,[38] and Vargas (2013) gives some acknowledgement of how blaming fosters agency (and thus implicitly fosters competence), a complete account of blame's sensitising ability must explicitly distinguish its effects on both motivation and competence. When directed at beings with a norm-psychology like ours, blame produces *both* competence and internalisation of norms. This explains why it can be somewhat tempting to focus more on one dimension over the other: since they develop mostly concurrently, giving a role to one basically allows the other's effects to come along in one's theory for free. But there are two distinct roles here, and the story of how blame sensitises us to norms is clearly one that occurs over many years and at different time-scales, which we as theorists need to ourselves be sensitive to.

Sensitisation is the final step in our story of blame's nature. Once an agent is sufficiently competent with a norm, and motivated to exercise that competence, they can then be relied upon. They are someone who shares our commitments, and whom we don't need to continually worry will take advantage of us.

6 Blame Theorising

Many people frequently misunderstand claims about costly signals, so I need to address some objections head-on. One common objection is that some blame doesn't seem like a *costly* signal: we can imagine cases where blame is private, directed at the dead who can't retaliate, or may even earn us certain benefits. Another is that some blame doesn't seem like a costly *signal*: particularly if we blame privately, or express

[38] Fricker seems to think that proper recognition of the moral reasons just does motivate us to act on them. McGeer (2019) later analogises learning norm-competence to learning other skills that take many years, but most of her focus is on interactions between agents who are already generally competent with our blaming practices, and who can 'negotiate' (p. 262) what our moral standards should be. McGeer (2013) also recognises that signalling and 'regulative' responses contribute to these interactions, but her goal is to show how our evolved responses were built upon and 'recalibrated' by our mentalizing capacities, being used to "develop the moral understanding of all parties in a normative dispute" (p. 175) by communicating to the wrongdoer. But as Reis-Dennis (2019) points out, given we can also communicate our sincerity or the seriousness with which we regard an offence without being angry, this account misses what makes angry blame *distinctive*: its ability to communicate our "willingness to fight" (p. 456), which I would identify as 'willingness to sanction'. Rather than focusing on individual attitudes, particular wrongs, and the goal of rectifying attitudes to help scaffold others', we gain a more instructive picture of blame by instead thinking about our goal of on-going cooperation, the importance of norm internalisation, and agents' commitments. An additional benefit of the latter framing is that it squares nicely with the recent work on hypocritical blame. Several philosophers have noted (entirely independently of Shoemaker and Vargas's account) that *commitments* to seem to be the key factor determining when blame counts as hypocritical (Friedman, 2013; Rossi, 2018, Piovarchy, 2023a, b; Isserow, 2022; Todd, 2019), and I believe this is not a coincidence. Blame's ability to signal commitments is precisely what explains why hypocrites lack the standing to blame: hypocritical blame is a form of dishonest signalling (see Piovarchy, forthcoming; Jeffers & Shaefer, 2025).

it but no-one sees. A third is that sanctioning isn't always our motivating reason for blaming. People blame for all kinds of reasons. And finally, it seems agents can blame without an increased disposition to sanction.

These kinds of worries—I need to be crystal clear on this point—*are not persuasive objections*. The claim 'blame is a costly signal of commitment to a norm' does *not* mean that for any given instance of blame, *that* token blame was costly for *that* token agent to express or feel, any more than 'engineering degrees are costly signals of engineering ability' entail that each and every person with an engineering degree had to incur high costs to get their degree.

What it does mean is this is the *type* of signal which *would* be *sufficiently* costly for the *types* of beings without the quality to send to affect their behaviour to a degree which enables receivers to *separate* types of senders with sufficient reliability. All that is needed is for a *correlation* to develop in sender *types*, which receivers notice and condition their response on. *This* claim is *perfectly* compatible with many beings without the quality emitting the signal.[39] It is even compatible with many beings without the quality emitting the signal and not incurring costs in doing so! It is also compatible with many beings blaming for all manner of reasons (just as one can get an engineering degree without desiring to be recognised as being good at engineering). Note that each of these objections results from the very error I identified at the beginning of this paper: trying to understand blame by taking an individual blamer as one's focal point. This is why I am emphasising the importance of avoiding falling into this trap: *individual cases are not even the type of thing that would show the claim is mistaken*. Such deep and persistent misunderstandings are why I've provided so much detail on the behaviour of agents in mixed-motive public goods games, our psychologies as agents in the situations modelled by those games, and exegesis on earlier arguments that overlap with this story.

I believe that failure to consider the broader context of blaming have lead much theorising astray. The moral responsibility literature used to recognise a much closer connection between sanctioning behaviours and our moral accountability practices, and our interest in such behaviours arguably motivates much of our original interest in questions about the latter. It seems that when most people become interested in questions about moral responsibility, what it means to hold people accountable, and when it is appropriate to do so, for instance, they would not predict that the answers to such questions would be exhausted by a sufficiently detailed account of when certain emotions and attitudes are fitting to experience, with questions of when it is appropriate to treat people in certain ways being some other, independent topic. To put it another way, if we insist that e.g. a signal of commitment is all that blame really is, we have to accept that much of the interest in questions like 'When are agents morally responsible for wrongdoing *qua* apt targets of blame?' should be understood as asking 'When are agents morally responsible *qua* apt targets of signal expression and a willingness to police norms?', with questions of when it is appropriate to enforce said norms being entirely distinct. Such a bifurcation seems to answer the question

[39] This is well-understood by biologists and economists who use costly signalling in their theories, and should not be considered a novel proposal requiring more defence. Still, matters are, in fact, considerably more complicated than space permits; see Brusse (2019) for a helpful overview.

of what blame is at the expense of reducing its bearing on what interested us in it in the first place.[40]

These interests favour considering blaming attitudes, signalled commitments, and enforcing behaviour as all part of blame proper. I have already noted how taking blaming attitudes and emotions to be more explanatorily basic than their expressions can be misleading, since part of what *makes* them costly signals are the behaviours they dispose us towards. I am not sure that our folk concept of 'blame' (certainly 'blaming') cleanly distinguishes between blaming attitudes/emotions and the behaviours that express such attitudes in the way that many philosophers would like, and this is plausibly because our interest in questions of accountability and how to treat people influence our conceptual usage.

My contention is that a more complete picture of blame requires understanding the interlocking nature of norms, norm-following, norm stability, and our norm-psychology, as it is from these factors that our moral accountability practices emerge. To understand what blame is, we understand what it does, which requires understanding what problems it solves, *and this requires understanding what kinds of beings we are and how norms are maintained in communities*. Trying to understand blame's nature by focusing on individual blamers alone leaves out too many of these factors. The moral community is not merely a set of individual blamers, multiplied, nor merely a black box from which our standards or collective demands emerge. It is made of agents who need to *coordinate, cooperate*, and *work together* to solve *very particular* kinds of problems.

Some readers may worry that having three distinct effects is theoretically untidy. But it is not, for several reasons. First, these effects are united by a common function: upholding norm-following. They have *not* been chosen merely as a means of salvaging a conceptually messy and contextually variant folk concept. Nor have they been selected in an ad hoc fashion to simply save one's pet theory against objections. Nor is this simply unprincipled pluralism. They are all effects that philosophers and scientists of many stripes have observed are tightly involved in blaming responses. Just as a heart's function is pumping blood, but it does this *by* contracting muscles and raising systolic blood pressure, which transports blood and results in oxygenated tissue, blame's function is upholding norms, but it does this by signalling and sanctioning, which stabilises norm-following and results in sensitised agents.

Second, each effect plays a distinct role, which we do not achieve stable norm-following without. Agents will not comply if they are incompetent, even if they are highly motivated to comply and know others are too. Nor will they comply if they are competent but unwilling, and several factors are relevant to agents' motivations. There will always be some agents who are tempted to defect out of self-interest, who thus need to be deterred with sanctions. And—as the existence of co-ordination failures that are in everyone's interests to overcome demonstrate—in many cases

[40] Similarly, Shoemaker's (2024) sharp distinction between what makes agents fitting targets of (private) blaming attitudes or emotions, and when it is fair to sanction according to the rules of our various interpersonal games (which he thinks does not depend on blameworthiness at all), seems to questionably sever the thing we are interested in: when and how we are justified in treating people by blaming them. More simply, it seems that blameworthy-making features of agents frequently change what kinds of sanctions are fitting (cf. McKenna, 2024, p. 186).

agents won't comply even if they are competent and care about the norm, because they may worry no-one else will. Agents often operate in noisy environments with imperfect information and enforcement mechanisms, some make accidents, and different communities have different norms. Uncertainty continually threatens to reduce norm-following, so finding a way to make agents' motivations robust enough to not need complete certainty before cooperating helps them avoid a whole range of costs. Upholding norms can be valuable, but keeping it *stable* requires we solve these problems.

Different philosophers writing on blame have noticed that blame produces these effects, but they've each opted to focus on one effect in their accounts, occasionally arguing the others are useful but contingent. Insofar as we're interested in a complete understanding of blame's nature, however, we ought to opt for a more (rather than less) detailed map of blaming attitudes, emotions, behaviours, and their effects.

This account weaves together components from several theories in a way that preserves much of what was appealing about each. But an important benefit is that it doesn't need to explain away any particular effect as merely 'contingent' because it shows there are principled connections between all three of them, and the underlying causal story is very simple. Agents have a commitment to a norm, which produces a range of dispositions that are triggered in response to norm violations, all of which express concern for the norm and an unwillingness to tolerate violations (signalling) and are characteristically accompanied by motivation to actually enforce the norm (sanctioning). These reactions are what we commonly identify as 'blaming'. When collectively practiced, they provide assurance and deterrence, and, over time, produce internalisation and competence (sensitising), as (being norm-sensitive beings) we come to care about the norm and learn how to comply with it.

Getting this story off the ground requires very minimal, uncontroversial assumptions. All that one needs to grant is that humans have a norm-psychology and use others as a guide for how to behave, that they don't like sanctions, that information, feedback and practice is useful for developing competence, and that blaming is able to signal commitments and sanction its targets. With these minimal assumptions, and an understanding of the ways that that norm-following can collapse, the three-pronged story of blame's nature developed above falls right out. It successfully accounts for the considerable heterogeneity in phenomena that are often identified as constituting blame, it gives a more comprehensive account of the role of blaming in our moral practices, it is naturalistic and compatible with our best theories of human psychology and biology, and it manages to synthesise features from a variety of existing, attractive accounts of blame into a single, unified story.

7 Conclusion

Blame is a heterogeneous, multi-purpose phenomenon. Philosophers have spent much time and effort trying to work out what its core features are, and what its main point is. The difficulty in creating consensus about blame stems from the fact that it has multiple effects, which can only be noticed by switching our focus from blamer-blamed dyads to norm-followers in communities. Blaming communicates important

information. It treats its target negatively. And it produces understanding, motivation, and competence. In short, it signals, sanctions, and sensitises. In doing so, it plays a very important but under-appreciated role maintaining norms, many of which are very valuable things to uphold. Keeping an eye on how blame occurs not just between perpetrators and blamers, but between perpetrators and communities, helps us recover a more complete understanding of blame's nature, and (to appropriate Strawson, 1962) a deeper "sense of what we mean, i.e. of all we mean, when, speaking the language of" our moral accountability practices (p. 24).

Acknowledgements For helpful discussion, feedback, or comments, I am grateful to Mark Colyvan, Mark Alfano, Hannah Tierney, and audiences at UNDA, Pardubice, ANU, Macquarie, and The University of Sydney.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Work on this paper was partly funded by grant DE260101744 from The Australian Research Council.

Declarations

Competing interests None.

References

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. (2014). Norms, conventions, and the power of expectations. In N. Cartwright & E. Montuschi (Eds.), *The philosophy of social science: A new introduction* (pp. 208–229). Oxford University Press.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press USA.

Bird, R. B., & Smith, E. A. (2005). Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology*, *46*(2), 221–248.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. Oxford University Press.

Brennan, G., & Pettit, P. (2005). *The economy of esteem: An essay on civil and political society*. Oxford University Press.

Brusse, C. (2019). Signaling theories of religion: Models and explanation. *Religion, Brain & Behavior*, *10*(3), 272–291.

Carlsson, A. (2017). Blameworthiness as deserved guilt. *The Journal of Ethics*, *21*(1), 89–115.

Caruso, G. D. (2021). *Rejecting retributivism: Free will, punishment, and criminal justice*. Cambridge University Press.

Chudek, M., & Henrich, J. (2011). Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, *15*(5), 218–226.

Chudek, M., Zhao, W., & Henrich, J. (2013). Culture-gene coevolution, large-scale cooperation, and the shaping of human social psychology. In K. Sterelny, R. Joyce, B. Calcott, & B. Fraser (Eds.), *Cooperation and its evolution*. MIT Press.
Coates, J. D., & Tognazzini, N. A. (2013). *Blame: Its nature and norms*. Oxford University Press.
Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69.
Davis, T., Hennes, E. P., & Raymond, L. (2018). Cultural evolution of normative motivations for sustainable behaviour. *Nature Sustainability*, *1*(5), 218–224.
Dixit, A. K., & Skeath, S. (2015). *Games of strategy* (3rd ed.). W. W. Norton & Company.
Driver, J. (2016). Private blame. *Criminal Law and Philosophy*, *10*(2), 215–220.
Edwards, C. P. (1987). Culture and the construction of moral values: A comparative ethnography of moral encounters in two cultural settings. In J. Kagan, & S. Lamb (Eds.), *The emergence of morality in young children*. University of Chicago Press.
Fehr, E., & Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994.
Fitzgibbon, C. D., & Fanshawe, J. H. (1988). Stotting in Thomson's gazelles: an honest signal of condition. *Behavioral Ecology and Sociobiology, 23*(2), 69–74.
Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. WW Norton & Co.
Fraser, B. (2012). Costly signalling theories: Beyond the handicap principle. *Biology and Philosophy*, *27*(2), 263–278.
Fricker, M. (2016). What's the point of blame? A paradigm based explanation. *Noûs*, *50*(1), 165–183.
Friedman, M. (2013). How to blame people responsibly. *Journal of Value Inquiry*, *47*(3), 271–284.
Graham, P. A. (2014). A sketch of a theory of moral blameworthiness. *Philosophy and Phenomenological Research*, *88*(2), 388–409.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.
Hieronymi, P. (2004). The force and fairness of blame. *Philosophical Perspectives*, *18*(1), 115–148.
Hieronymi, P. (2019). 'I'll Bet You Think This Blame is About You'. In D. Justin Coates, & Neal A. Tognazzini (Eds.), Oxford studies in agency and responsibility volume 5: Themes from the Philosophy of Gary Watson, Oxford studies in agency and responsibility.
Isserow, J. (2022). Subjunctive hypocrisy. *Ergo, 9*(7), 172–199.
Jeffers, M., & Shaefer, A. (2025). The function of Hypocrisy norms. *Utilitas*, *37*(2), 123–140.
Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2635–2650. https://doi.org/10.1098/rstb.2010.0146.
King, M. (2014). Two faces of desert. *Philosophical Studies*, *169*(3), 401–424. https://doi.org/10.1007/s11098-013-0188-5.
Macnamara, C. (2013). Taking demands out of blame. In J. D. Coates, & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 141–161). Oxford University Press.
Mappes, J., & Alatalo, R. V. (1997). Batesian mimicry and signal accuracy. *Evolution*, *51*(6), 2050–2053.
McGeer, V. (2013). Civilizing blame. In D. J. Coates, & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 162–188). Oxford University Press.
McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, *18*(2), 259–281.
McGeer, V. (2019). Scaffolding agency: A proleptic account of the reactive attitudes. *European Journal of Philosophy*, *27*(2), 301–323.
McKenna, M. (2012). *Conversation & Responsibility*. Oxford University Press.
McKenna, M. (2016). Quality of will, private blame and conversation: Reply to driver, shoemaker, and vargas. *Criminal Law and Philosophy*, *10*(2), 243–263.
McKenna, M. (2024). *Responsibility and desert*. Oxford University Press.
Menges, L. (2021). The kind of blame skeptics should be skeptical about. *Canadian Journal of Philosophy*, *51*(6), 401–415.
Mill, J. S. (2015). *On liberty, utilitarianism, and other essays*. Oxford University Press.
Nelkin, D. K., & Pereboom, D. (2022). *The Oxford handbook of moral responsibility*. Oxford University Press.
Nowak, M., & Sigmund, K. (1993). Chaos and the evolution of cooperation. *Proceedings of the National Academy of Sciences*, *90*(11), 5091–5094.
Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.

Pereboom, D. (2021). *Wrongdoing and the moral emotions*. Oxford University Press.
Pinker, S. (2025). *When everyone knows that everyone knows…: Common knowledge and the science of harmony, hypocrisy and outrage*. Random House.
Piovarchy, A. (2023a). Sitationism, subjunctive Hypocrisy and standing to blame. *Inquiry*, *66*(4), 514–538.
Piovarchy, A. (2023b). Does being a 'Bad feminist' make me a hypocrite? Politics, commitment and moral consistency. *Philosophical Studies*, *180*, 3467–3488.
Piovarchy, A. (forthcoming). Hypocritical blame as dishonest signalling. *The Australasian Journal of Philosophy*.
Portmore, D. W. (2022). A comprehensive account of blame: Self-blame, non-moral blame, and blame for the non-voluntary. In A. B. Carlsson (Ed.), *Self-blame and moral responsibility* (pp. 48–76). Cambridge University Press.
Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, *44*(3), 875–881. https://doi.org/10.1037/0012-1649.44.3.875.
Reis-Dennis, S. (2019). Anger: Scary good. *Australasian Journal of Philosophy*, *97*(3), 451–464.
Rosen, G. (2003). Culpability and ignorance. *Proceedings of the Aristotelian Society*, *103*(1), 61–84.
Rossi, B. (2018). The commitment account of Hypocrisy. *Ethical Theory and Moral Practice*, *21*(3), 553–567.
Russell, P. (2017). Review of Michael McKenna, conversation and responsibility. *Philosophical Review*, *126*(2), 285–295.
Scanlon, T. M. (2008). *Moral dimensions: Permissibility, meaning, blame*. Belknap Press of Harvard University Press.
Scanlon, T. M. (2013). Interpreting blame. In J. D. Coates, & N. A. Tognazzini (Eds.), *Blame: Its nature and norms*. Oxford University Press.
Schlick, M. (1939). *Problems of ethics* (Vol. 36). Dover Publications.
Schmidt, M. F., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, *27*(10), 1360–1370.
Sher, G. (2005). *In praise of blame*. Oxford University Press.
Shoemaker, D. (2017). Response-dependent responsibility; or, a funny thing happened on the way to blame. *Philosophical Review*, *126*(4), 481–527.
Shoemaker, D. (2024). *The architecture of praise and blame*. Oxford University Press.
Shoemaker, D., & Vargas, M. (2021). Moral torch fishing: A signaling theory of blame. *Noûs*, *55*(3), 581–602.
Skyrms, B. (2003). *The stag hunt and the evolution of social structure*. Cambridge University Press.
Smart, J. J. C. (1961). Free will, praise and blame. *Mind*, *70*(279), 291–306.
Smith, A. M. (2013). Moral blame and moral protest. In J. D. Coates, & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (Vol. 27-48). Oxford University Press.
Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*(5), 495–518.
Sripada, C., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. P. Stich (Eds.), *The innate Mind, Volume, 2: Culture and cognition*. Oxford University Press.
Stich, S. (2018). The quest for the boundaries of morality. In *The Routledge handbook of moral epistemology* (pp. 15–37). Routledge.
Strabbing, J. T. (2019). Accountability and the thoughts in reactive attitudes. *Philosophical Studies*, *176*(12), 3121–3140.
Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 1–25.
Sugden, R. (2010). *Is there a distinction between morality and convention?*. Nomos Verlagsgesellschaft mbH & Co. KG.
Talbert, M. (2012). Moral competence, moral blame, and protest. *The Journal of Ethics*, *16*(1), 89–109.
Thrasher, J., & Handfield, T. (2018). Honor and violence: An account of feuds, duels, and honor killings. *Human Nature*, *29*, 371–389.
Todd, P. (2019). A unified account of the moral standing to blame. *Noûs*, *53*(2), 347–374. https://doi.org/10.1111/nous.12215.
Tomasello, M. (2015). *A natural history of human morality*. Harvard University Press.
Vargas, M. R. (2013). *Building better beings: A theory of moral Responsibility*. Oxford University Press.
Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.
Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227–248.

Watson, G. (Ed.). (1987). *Responsibility and the limits of evil: Variations on a Strawsonian theme*. Cornell University Press.
Wolf, S. (2011). Blame, Italian style. In *Reasons and recognition: Essays on the philosophy of T. M. Scanlon* (pp. 332–346). Oxford University Press.
Young, H. P. (1998). Social norms and economic welfare. *European Economic Review*, *42*(3–5), 821–830.
Zahavi, A., & Zahavi, A. (1999). *The handicap principle: A missing piece of Darwin's puzzle*. Oxford University Press.