Moral Torch Fishing:
A Signaling Theory of Blame

David Shoemaker & Manuel Vargas

Forthcoming in *Noûs*

ABSTRACT: It is notable that all of the leading theories of blame have to employ ungainly fixes to deflect one or more apparent counterexamples. What these theories share is a content-based theory of blame's nature. Such approaches overlook or ignore blame's core unifying feature, namely, its *function*, which is to signal the blamer's commitment to a set of norms. In this paper, we present the problems with the extant theories and then explain what signaling is, how it functions in blame, why appealing to it resolves the problems in other theories, what the signaling function implies for a wider range of gray-area cases, and what the larger significance is of blame's core function for interpersonal interactions in a variety of (non-moral) normative domains.

1. Torch Fishing

Ifaluk is a coral atoll in Micronesia. For its inhabitants, who work for subsistence, their primary source of protein comes from fish. Only men do the fishing. There are several different types of fishing they engage in, most of which aim to capture yellowfin tuna, but during the trade wind season, some men engage in torch fishing, a highly ritualized form of fishing that takes place at night. There are two stages. In the first, the men use torches they have hand-woven out of layers of coconut fronds to attract flying fish into a small net. Then they use those fish as bait in deep-sea fishing for dogtooth tuna. The preparations for torch fishing take several weeks. The men who will participate have to collect and dry many coconut fronds and fashion them into torches. Many men also make the hand nets to catch the flying fish. After engaging in regular fishing over the course of the day, they have to fish much of the night during this cycle of the moon. The men who participate sleep very little during this period, and some not at all. It is rigorous, difficult work. And for all that work, the amount of fish gathered is far less than that gathered by the other everyday methods. Indeed, the whole process tends to involve a net caloric loss.

So why do they do it? No one has to engage in the ritual, and many men don't. Their ritualized efforts to catch the dogtooth tuna seem on their face quite irrational; everyday fishing styles are much more efficient. But there is a way to view what they are doing that makes perfect sense of the practice. They are benefiting in the long run by *signaling* something with their actions, conveying information to the women and other men who gather on the beach to watch the ritual. What they are signaling, it is thought, is that they have an excellent work ethic and will be good providers. This increases their odds of getting into beneficial marriages or achieving high community status (Sosis 2001). So the significance of torch fishing is less about the fish and more about what the fishing signals.

In this paper, we will argue that blame is essentially moral torch fishing. It is a costly response to norm violations defined most fundamentally not by any particular content—e.g., a mental state or activity—but by a *function*, namely, the signaling of the blamer's commitments, including a commitment to the enforcement of those commitments. This is the remarkably underappreciated unifying feature of blame. To make our case, we will briefly motivate the project by showing the fancy dancing that content-based theories of blame have to engage in in order to account for a variety of outlier cases. We will then diagnose the problem and advance our own functional solution, the signaling theory of blame, which builds on and extends the Costly Signaling Theory as it is understood and applied in a variety of other disciplines. Once we have provided the details of our theory, we will explain, not only how it deals easily with the outlier cases, but also how it has additional explanatory payoffs with regard to a number of features of our interpersonal lives. Indeed, thinking about blame in terms of signaling enables us to see that blame itself is somewhat less distinctive than it is ordinarily thought to be; rather, it is a species of a more fundamental and fascinating inter- (and intra-) personal dynamic. We will conclude by providing some suggestions for how to prove us wrong.

2. Blame's data points

Nearly all the leading theories of blame agree about at least this much: *blame is a response to a person in light of his or her perceived norm violation, where the blamer takes that violated norm seriously*.[1] Theories of blame in the philosophical literature have been proposals for fleshing out the precise nature of that response. What makes all the leading theories of a piece is that they seek to identify the blaming response with its *content*, namely, an attitude or mental state (Arpaly and Schroeder 2014: 159-160 say this explicitly; see also Wolf 2011: 334). We

[1] What does it mean to "take seriously" the norm? We leave this as a rough-and-ready intuitive notion for now, one we will later explicate.

think there is a problem inherent in any such a strategy that is resolved only by taking a very different approach to understanding blame. In this section, after listing the data points that theories of blame aim to account for, we will briefly discuss versions of the four leading theories of blame in order to demonstrate the difficulties they each have—in different respects—in accounting for some items on the data list. We will then turn to diagnosing the problem and offering our alternative solution.

What must a theory of blame account for? There are many data points. Here are just some of them:[2]

- *Blame involves more than a mere belief that the norm violator has acted wrongly* (Sher 2006: 6): Believing that you slighted me[3], say, is not yet blame. I may take myself to deserve that treatment, for instance, or I may be delighted that you slighted me, for in so doing I won a bet (cf. McKenna 2012: 22).
- *Directed, dyadic overt blame*: Most theorists take this to be the paradigm form of blame.[4] It involves two parties, the offender and the offended, wherein the latter blames the former directly and openly.
- *Non-directed overt blame*: Sometimes when slighted I do not blame the slighting agent face-to-face, perhaps because it is my boss or a dangerous person. But I may well blame him to friends, family, co-workers, or bystanders.
- *Private blame*: Sometimes I blame an offender only to myself, without expressing it in any way to anyone else.
- *Blaming the dead*: Even when there is no chance whatsoever that my blame can be communicated to the offender, I may still blame him or her.
- *Self-blame*: Sometimes the blamer and the blamed are one and the same individual.
- *Dispassionate blame:* Sometimes we may be utterly without emotion in blaming someone. Perhaps a mother blames her repeatedly wayward son with nothing but exhaustion.
- *Hypocritical blame*: Perhaps I am a serial promise-breaker, but when you break a promise to me, I blame you for it. There is something "off" about such blame (even though it is blame), but what?
- *Hypothetical blame*: Consider a couple watching the movie *Force Majeure*, about a husband who abandons his wife and children as he runs in terror from a threatening avalanche. After watching this scene, one partner angrily elbows the other and says, "That's something *you* would do."[5]

Our argument in what follows is that all the leading theories of blame have prima facie difficulty accounting for some of the items on the list, that is, they have to engage in what we call *fancy dancing* in order to do so. Our ambition is to highlight the challenges facing all the going accounts. We acknowledge that every account has things they can say about these issues. Our point is not that they cannot address them in any fashion at all; rather, we hope to highlight that there is a reason why they are forced to engage in fancy dancing in the first place, and identifying what it is should motivate us to adopt a radically different approach to explicating blame's nature.

---

[2] For a more complete list, see Shoemaker 2013: 101.

[3] Even though we are a we, we will use "I" and "me" in articulating some examples, as this is a more natural and familiar locution.

[4] McKenna is one of the many who explains in detail why what he calls "overt" blame is the more fundamental thing to be explained (see McKenna 2012: 174-78; and 2013). Arpaly 2006 (9) favors private blame as more fundamental, as do Carlsson (2017) and Portmore (Forthcoming), but as of now they remain in the minority.

[5] Some readers may deny that this last entry is "really" blame. One of our many aims in this paper is to cast doubt on that conviction.

3. Fancy Dancing

The most influential approach to blame has been the *reactive attitudes theory*, according to which blame fundamentally consists in a range of emotional responses, typically forms of anger. This view has its roots in P.F. Strawson's "Freedom and Resentment" (Strawson 2003), in which he discusses a variety of reactive attitudes, focusing primarily on resentment and indignation.[6] As R. Jay Wallace puts it, it would be "strange to suppose that one might blame another person without feeling an attitude of indignation or resentment toward the person..." (Wallace 1994: 75).

But it has not seemed so strange to some. Indeed, there are plenty of prima facie false negatives and false positives to this view. Regarding the former, George Sher notes that we may "feel no hostility toward the loved one whom we blame for failing to tell a sensitive acquaintance a hard truth, the criminal whom we blame for a burglary we read about in the newspaper, or the historical figure whom we blame for the misdeeds he performed long ago" (Sher 2006: 88.). Similarly, Angela Smith notes that I may respond to your wrongdoing "by dispassionately 'unfriending' [you] on [my] Facebook page... or by simply refusing to trust [you] anymore, and these too should qualify as forms of blame" (Smith 2013: 32). We may also engage in dispassionate blame of persons for various poor character traits, as when I merely express my judgment that you are a cruel or cowardly person (*pace* Watson 2004: 266).[7] Anger and its ilk are unnecessary for blame.

There are also false positives, cases in which the specified reactive attitudes are in place without blame. We can, for instance, be indignant about the maladies of aging (the loss of youthful reaction time and recall), or resentful for being alive.

It was in part to avoid similar worries that T.M. Scanlon developed the *relationship-modification theory* of blame (Scanlon 2008: Ch. 4, esp. 127-8). On this account, blame consists in two mental states or activities: (a) a judgment that someone has done something reflecting "attitudes toward others that impairs the relations that others can have with him or her" (Scanlon 2008: 128); and (b) a modification of one's own attitudes, intentions, or dispositions in a way appropriate to that relationship impairment. The latter might include feeling various reactive attitudes, of course, but it also might not (I might just omit my ordinary friendly greeting).

The flexibility of this theory allows it to capture the cases the reactive attitudes theory has trouble with, but it has trouble itself with others. Susan Wolf, in response to Scanlon, provides a false negative by pointing to cases of blame between close family members (what she calls "Blame, Italian Style"), where despite the screaming and remonstration, no relationship modification in the least occurs, despite judgments of wrongdoing and blameworthiness by the blamers (Wolf 2011: 334).

There are prima facie false positives for the relationship-modification theory as well. Angela Smith provides the example of a mother whose son has committed a horrible crime. She judges that he is blameworthy, as he has impaired many relationships, but perhaps to compensate for the hate others will rain down upon him, she amplifies her love and commitment to him, yet also modifies (i.e., significantly lowers) her expectations that he'll one day be a great business success. These attitudinal modifications, plausibly appropriate responses to a mother's judgment of her son's blameworthiness, are nevertheless not blame (Smith 2013: 38).

---

[6] Strawson doesn't use the term "blame" for these reactive emotions, but as Scanlon points out, "this identification is a natural application of his analysis" (Scanlon 2008: 224, n. 6). Others taking up this cause include Wallace 1994, 2011; Wolf 2011; McKenna 2012; Vargas 2013; Bennett 2013; and, at least for accountability-responsibility, Shoemaker 2015: Ch. 3.

[7] Perhaps, though, these are *inapt* expressions? It doesn't matter. A theory about blame's nature has to capture what's common to both apt and inapt versions of it.

To sidestep these problems, Smith (along with some others) advances our third theory, the *protest theory* of blame. On this view, your blame consists in those attitudinal or behavioral responses that repudiate or take a stand against—protest—someone's immoral treatment of you. Such protest registers and challenges the fact that you didn't deserve to be treated in this way, and it aims to get the offender and others in the moral community to acknowledge what the offender did, and moral threat implicit in his conduct (Smith 2013: 41-47; see also Hieronymi 2004, and Talbert 2012). This view allows for a wider range of attitudes to count as blame than what is countenanced by the reactive attitudes theory, retaining the advantage of the relationship-modification theory, but in adding the protest element, it avoids the false positives attached to that theory (Smith 2013: 40, 45).[8]

However, if blame is fundamentally about registering objections to and seeking acknowledgment from others, then *private blame*—a quiet resentment or indignation, say—looks like it cannot be blame, so the view has a prima facie false negative.[9]

There are false positives too. When Sister Helen Prejean stands outside a prison with a lit candle, she protests the (what she takes to be wrongful) execution of a condemned person taking place, but she does not (necessarily) blame anyone. Less weighty but more familiar examples involve parents protesting the naughty behavior of their children without blaming them.

Protest alone also doesn't match up very well with what are often taken to be blame's aims, which are to get the wrongdoer in line. It has thus made sense to many to restrict the relevant blaming content to attitudes or activities that deliver various injunctions, seek uptake, and demand acknowledgment of wrongdoing and/or various forms of repair from the protested party. This would make *communication* fundamental to blame (with protest being just one form of it), and so yields our fourth theory of blame's core mental state or activity, what we will creatively call the *communicative theory* of blame. Many have held a version of such a view.[10] For all of these authors, certain (e.g., wrongful) actions and attitudes render blaming responses appropriate, where these attitudes have a built-in communicative aim, e.g., to express a moral reminder or demand to the wrongdoer (Watson 1987; Darwall 2006; Shoemaker 2007), or to invite the wrongdoer to respond with acknowledgment, apology, reparation, or moral improvement (Vargas 2013; Macnamara 2015; Shoemaker 2015).

On its face, the communicative theory runs into some of the same problems as the protest theory. The problem of private blame still lurks, and people may communicate moral demands or invitations without blame

---

[8] Coates and Tognazzini classify Smith's view as a functional account of blame, insofar as she identifies blame with protest (what it does), as opposed to identifying it with any particular mental state (Coates and Tognazzini 2013a: 16). If it is a functional account, we think it identifies the wrong function. However, we doubt it is a functional account. Smith explicitly builds on and refines Scanlon's (attitudinal) theory by restricting the relevant relationship-modifying attitudes to those that protest the blamed agent's relationship-impairment (Smith 2013: 39). But this obviously a functional account so much as a conjunctive mental state account. On her account, a relationship-modifying mental state isn't blame unless it's accompanied by *another* mental state, namely, an *attitude of repudiation*, one which is roughly constituted by an intention to register an objection to the wrongdoing and "implicitly seeks some kind of moral acknowledgment on the part of the blameworthy agent and/or on the part of others in the moral community" (Smith 2013: 43). So ceasing interaction with a friend counts as blame only to the extent it is accompanied by this further protesting attitude, which is itself just another mental state.

[9] Smith deals explicitly with this sort of case, noting that "we can also protest ill treatment privately through the modification of other attitudes, intentions, and expectations" (Smith 2013: 44). We believe this stretches the notion of what counts as "protest" too far, however.

[10] Advocates or sympathizers include Gary Watson's reconstructed Strawson (Watson 1987), Watson himself (2010), Stephen Darwall (2006), David Shoemaker (2007), Michael McKenna (2012; 2013), and Coleen Macnamara (2013; 2015).

(e.g. Sister Helen). But the communicative theory has resources the protest theory may not have to deal with these cases. One might, for instance, say that the face-to-face nature of paradigm blame informs the nature of private blame, so that my secret response to an offender still counts as blame given that it involves attitudes that, *were* they in fact expressed to the offender, "would play a role in a kind of conversational [communicative] exchange with the one blamed" (McKenna 2012: 177).

Regarding the false positives that hound the protest theory, the communicative blame theorist can—and typically does—maintain that what's missing in these cases is a certain kind of moral anger,[11] which we only understand as such in light of what its public manifestation would be. So what makes some private emotional experience resentment "is precisely an appreciation of the criterial indicators that would be manifested in a public display of *that* emotion rather than some other emotion…" (McKenna 2012: 69; emphasis in original). What's missing in the Sister Helen case is the *resentment* whose manifestation would play this public role.

Notice, though, that in addressing the problems of the protest theory, the communicative theory has fancily danced its way right back to the original reactive attitudes theory. But then more fancy dancing is required to account for blame without reactive attitudes (relationship modification?), which puts us right back on the wheel of misfortune.[12]

Each of the leading theories offers a different specific mental state or activity as blame's core, but each on its own admits of prima facie false negatives and positives. What we haven't yet considered, though, is a disjunctive theory of blame's necessary condition, and a conjunctive theory of blame's sufficient condition. It might be, for example, that you can't have blame without at least *one of* the four mental contents (reactive attitude, relationship-impairment, protest, or communicated demands/invitations). And it also seems that where one has the entire *set* of listed mental states or activities, you're guaranteed to have blame.

We are inclined to agree with both points. But we also think that taking this permissive route gives the explanatory game away. For we want to know what *explains* why these mental states and activities, but not others, are together the essential constituents of blame. What, in other words, do feeling resentment, modifying relationships, protesting wrongdoing, and communicating moral demands *have in common*, such that each in their own way may constitute blame?

We believe that the content-based approach is simply inadequate to this explanatory task. Our blaming practices constitute a recognizably organized system, though. What we want to know is how each of the four mental states or activities contributes to the capacity of that organized system to be what it is (a blaming system). But this is just to say that what we need to look for is a *functional explanation* of the system, that is, a sensible explanation of what all this blame is *for* (Cummins 1983; Couch 2017). It is, we believe, most fundamentally for the sending of a costly signal.[13]

---

[11] See, e.g., Watson 1987; Wallace 1994; Nichols 2007; McGeer 2013; McKenna 2013: 124-5; Shoemaker 2015: Ch. 3.

[12] One further theory worth noting here is the proposal advanced by Sher (2006), according to which blame consists in a belief-desire pair: a belief that someone acted badly or had a bad character plus a desire that she not have acted badly or been that way. Caring about morality tends to trigger different kinds of dispositions to respond that have this belief-desire pair as their source, including various reactive attitudes, protests, relationship-modifications, and more (Sher 2006: 112, 138). Many have already noted the serious challenges such a view faces (see McKenna and Vadakin 2008, Smith 2013; McGeer 2013; and Franklin 2013). For our part, we think it explains too much, generating the sort of false positives already discussed, e.g., Smith's mother-son case. Or consider a case meeting Sher's conditions in which I respond only with horror (not blame) to a Syrian dictator's using chemical weapons on his own citizens.

[13] This is akin to the psychological functional approach laid out by McGeer 2013 and suggested by Nichols 2007. Among the things that differentiate our approach from theirs is that (a) we think blame incorporates much more than mere anger (or reactive attitudes generally), which is their only focus; (b) our approach is not necessarily about or connected to sanctions or punishment, something

4. Costly Signaling Theory and Blame

In many instances, there seems to be no net payoff for blame. It often costs a lot, including emotional equanimity, time, energy, self-control, and self-governance. It can even cost the blamer dear friends and lovers. And for what? To make people feel bad for something they can't go back and change, and that may even spark resentment on *their* part?[14] How could this make any rational sense?

Costly Signaling Theory (CST) explains how. CST has its roots in seminal works by Thorstein Veblen (1994/1899) and Marcel Mauss (1924), each of which detailed costly communications of a certain kind of status. The theory was essentially named and developed explicitly as an economic theory by Michael Spence (1973 and 2002), focusing on how information is communicated between individuals who have asymmetrical information in some market. So, for example, prospective employees signal their high qualifications to employers (for whom it's difficult to perceive such things directly) by their educational pedigree. This explanatory framework has been adopted since in a variety of disciplines, including organizational management, social psychology, anthropology, evolutionary biology and ecology, decision theory, legal theory, and religious studies.

Within evolutionary game theory, a costly signal could arise and become part of an evolutionarily stable system when the following conditions are in place:

1. Some members of a group have a quality that is difficult to perceive directly but to which a reliable signal could attach.
2. There are some members (observers) who would stand to gain from gleaning accurate information about that quality.
3. Signalers and observers have a conflict of interest, so that signalers who could successfully deceive observers about the quality would be benefited at the observers' expense.
4. The cost of the signal must have some benefit to the signaler (Bird and Smith 2005: 224).[15]

When a costly signal becomes part of some stable system, it will be one in which it has observers, it is hard to fake (otherwise it would be too easily imitated), it delivers accurate information to the observers, and it benefits the signaler (Bird, Smith, and Bird 2001). This, we maintain, is exactly how blame works in our interpersonal practices.[16]

---

they hang their hats on; (c) they are more interested in a story about blame drawn from the *original* and/or originally *stabilizing* function of the blaming response in evolutionary history, rather than a story that incorporates the *contemporary* function of the blaming response and takes seriously our cultural and ecological history in so doing; and (d) the signaling device we think is found in blame is also found in a number of other not-necessarily-blaming-but-nevertheless-blame-like interpersonal practices. Nevertheless, we certainly take both McGeer and Nichols to be forerunners and allies in the general cause. Other allies likely include John Doris (2015: 149-150), who has an insightful account of the signaling role of *apologies* in interpersonal exchanges; Christopher Bennett (2013), who discusses the symbolic function of expressing disapproval; and Thrasher and Handfield (2018), who detail the signaling function of honor-based violence and killing. All of this attention to signaling (and functional explanations more generally) warms our heart.

[14] This is the feature of blame that Nussbaum (2016) seems to find deeply incoherent, and is one basis of her rejection of it. For a response to Nussbaum on this point, see Shoemaker (forthcoming).

[15] What counts as a *cost* in signaling systems is a matter of some controversy (see Zahavi and Zahavi 1997; Maynard-Smith and Harper 2003; Searcy and Nowicki 2005; and Fraser 2012). We are sympathetic to a rather inclusive view, especially for human signaling systems (following Fraser 2012).

[16] We take the evolutionary game story to be a good model of how costly signaling works, but we are agnostic about whether blame *evolved* as a costly signaling function. Our proposal only requires that it *currently* functions in this way, quite apart from whatever its origins might have been.

So let's consider a paradigm case. You've made a promise to help me move that, when the time comes, you simply blow off. The next time I see you, I angrily protest what you did, communicating to you a demand for acknowledgment and apology, and letting you know that I won't help you the next time you need help. In other words, I blame you. All of this activity clearly conveys a lot of otherwise hard-to-read information to you, i.e., it signals to you loud and clear that it really matters to me that you not break such promises and that I'm willing to enforce promises made to me. This signal is hard to fake—hard enough that even professional actors can fail to be fully convincing in capturing the involved attitudes, facial expressions, and bodily cues.

There are many costs embedded in this particular signal. All sorts of unpleasant emotions have been stirred up in me. I have to invest time and energy in responding to you in this way, in carrying out or expressing the blame. I am also motivated to act in ways that involve risks, e.g., you, as the blamed agent, may respond in unexpected or nasty ways. I also risk the end or corrosion of our relationship. Blame occasionally costs one friends and loved ones.

But there are also substantial costs in my being a blaming *agent*, someone disposed to respond (almost automatically) as I just have. I have had to invest enormous social and psychological resources to maintain my membership in the blaming system. That is to say, I have had to keep up with what the incredibly wide-ranging and subtle interpersonal norms for interpersonal interaction are, for they regularly evolve, and I have also had to cultivate dispositions to respond to their violations in recognizable and normatively authorized ways. I can no longer, for instance, respond to your promise-breaking by slapping your face with my glove and "demanding satisfaction." Membership has its privileges, but the membership fees for Club Blame require not just memorization but (a) *internalization* of the norms and (b) attendant patterns of actual *enforcement* responses. We'll return to these crucial pieces in a moment.

Even when some (or even many) of these costs aren't implicated in any particular instance of blame (perhaps I just dispassionately reproach you and immediately move on), the reliability of what I nevertheless signal may depend on there being punishment awaiting for *me* were I to send a dishonest signal (Fraser 2012). Most of us tend to think of people whose blame we can see through as attempting to signal a commitment to norms that they don't actually have as busybodies or grandstanders (see, e.g., Tosi and Warmke 2016), and they are susceptible to (at least) reproach for their dishonesty. The threat of reproach or sanction also tends to keep would-be blamers in line for the *degree* of their blame. Blaming too much or too little also tends to invite negative responses, so honest signaling requires having refined blaming sensibilities, which itself incurs additional training and internalization costs.

Given all these costs, then, how could it possibly be rational to signal? We have already glimpsed the answer in the torch fishermen of Ifaluk. In enforcing the norms you violated in the way I do, I convey otherwise hard-to-read information to you about myself and what I value: The norm(s) you violated matter to me, and I'm not the type of person to let such violations pass. I may also be conveying an invitation for you to affirm your commitment to these norms as well. What I'm doing most fundamentally is signaling that I'm a member of a particular moral tribe, someone who cares about a set of norms and their breaches, someone who is disposed to police the norms, and more. That is to say, I am signaling what I am committed to, what practical reasons I take seriously, and *who I am as a practical agent,* which can include expectations about how I am to be treated.

We turn now to highlighting and developing the two crucial features of our account mentioned above: norm internalization and enforcement. These have particularly distinctive significance in the case of blame, and the psychology involved is richer and more complicated than what is involved in mere reputation management, as some simpler versions of Costly Signaling Theory may suggest. In the case of my blaming you, what makes my signal reliable is that my blaming reactions and their full nuance are difficult to fake unless I have

internalized the relevant norms and dispositions of enforcement response to their violation. Here, emotions often play a key role. The emotional reactions typically involved in blame are, as Robert Frank (1988: 4-7) labels them, *commitment devices*. They irruptively and urgently commit me to act in a way that can be contrary to my immediate (or obvious) interests, but in doing so they can nevertheless provide me with long-term advantages.[17] But there may also be many other *non*-emotional features of blame that are hard to fake, including facial expression, bodily posture, and tone of voice (e.g., sharpness, sarcasm).

The signal's reliability is thus underpinned not only by my internalization of the relevant norm(s), but also by my emotional commitment to it, manifested in a willingness to police breaches of those norms. That I am discernibly disposed to endorse the norm and to police its enforcement are the grounds for my trustworthy reputation. Having such a reputation tends to bring (in the long run) all sorts of goodies, including (at a minimum) a solution to the many everyday prisoner's-dilemma-type situations we find ourselves in. (Consider: my reliable and genuine anger at a given norm violation gives you some reason to think I can be trusted not to cheat you in a related situation where doing so would otherwise be to your disadvantage and my benefit.)

On content-based accounts, blame is a distinctive attitude or activity. The practice of blame—that is, blaming—is then understood in terms of where and how that attitude or activity occurs. Our proposal inverts that relationship: it is the function of blame—a signal about our normative commitments—that determines which attitudes and activities are instances of blame when they are. In the typical case, an agent's internalized norms and his or her disposition to enforce those norms are what generate the signal's reliability.

One can, of course, attempt to fake the appearance of having internalized the norms and of having the disposition to police the violations. Psychopaths, for example, typically endeavor to do so. However, given the way that norm internalization further shapes our blame sensibilities, and given the way our dispositions of response to norm violations are shaped and refined by a complex system of social and psychological calibration, agents who are not genuinely committed to these norms will not perceive or police the moral features of the world in the same recognizable way as those who have done so. They will tend to stand out, sometimes blatantly. Blame signaling is costly in both its maintenance and in its exercise. It is also beneficial in the longer term. At its core, blame is moral torch fishing.[18]

Although blame and punishment are distinct things—we can blame without punishing, as in the case of blame that is not expressed—it may be useful to say a bit about the relationship of our account to well-known theories of altruistic punishment (Fehr and Gächter 2002; Fowler 2005; Egas and Riedl 2008). Within behavioral economics, accounts of altruistic punishment arose in the context of studies of public goods games. In a public goods game, participants individually invest their resources into a public pot, and the total amount invested is multiplied and evenly distributed among the players. The highest *collective* payoffs are secured if everyone invests all their resources, but the highest *individual* payoff is secured by a player who doesn't invest but still collects an equal share of a public pot that others invested in.

Experimental findings show that in public goods games, contributing players are prepared to engage in costly punishment of free riders, even when they won't interact with those free riders again. Moreover, punishment enables stable and ongoing cooperation: where there is no punishment, cooperation (measured by both the number and size of individual contributions to the pot) drops; where there is punishment, rates of cooperation increase and stabilize at a relatively high level.

---

[17] Given his behavioral economics orientation, it is unsurprising that Frank thinks of commitment devices as providing *incentives* for agents (1988: 4). For our purposes, what matters is the idea of agents being *committed* or *bound* to some course of action.

[18] We'll return to this downstream, but here "moral" should be understood in a very broad way.

Altruistic punishment may look puzzling if one thinks human motivations are entirely selfish. However, the now-standard explanation of this phenomenon is relatively straightforward: a class of negative emotions spur punishment, and creatures with these emotions are committed to altruistic (read: costly) punishment, which in turn enables stable cooperation (Fehr and Gächter 2002), where the involved mechanism fits with going models of evolutionary psychology (Gintis 2003; Fowler 2005; Egas and Riedl 2008), at least under a range of conditions (Egas and Riedl 2008).

Our account and this account of altruistic punishment both allow that emotions can function as commitment devices, ensuring norm enforcement even when it is costly to the enforcer. However, our account of blame differs in a few important ways from the account of altruistic punishment.

First, recall blame's variegated nature. Where altruistic punishment is explained entirely by emotion-driven responses to free-riding, on our account blame is only sometimes an expression of emotion. The standard story of altruistic punishment fails to explain the kind of puzzle that blame presents to theorists. In contrast, a signaling story can explain why many instances of blame are not emotion-driven responses to free riding.

Second, it is worth noting that the standard story of altruistic punishment was introduced as a *competitor* to costly signaling accounts that attempted to explain punishment purely in terms of reputation management (Fehr and Gächter 2002: p. 137). In contrast, we think these stories are complementary: the basic mechanisms of the altruistic punishment story just are the kinds of things that help explain the conditions of the *reliability* of blame's signal, but the only way to capture the diversity of blame itself is to recognize the explanatory priority of the signaling function.

In the next section, we show how a signaling theory of blame can straightforwardly account for—and provide genuine insight into—the enormous complexity of everyday blame.

5. Signaling Blame: Its Features

Consider the following case:

> Tony likes to eat his lunch near the sidewalk and watch the women stroll by. Today he calls out to a young woman (Bev), "Hey, baby! Why don't you smile? Don't worry, I'm definitely smiling as I watch you walk away!" He then makes a series of kissing sounds. Sarah, walking directly behind Bev, stops and yells at Tony: "What is your problem? Why can't you leave women alone? Don't you understand how uncomfortable that is to hear, how *harassing* it is?"

Sarah obviously blames Tony. She is angry, communicates a demand, protests his behavior, and, to the extent they are in some kind of moral relationship, she alters that relationship with him thereby. But she is also, crucially, signaling. What does her signaling consist in, precisely?

- *Normative Competence*: The most important aspect of signaling is that it conveys lots of information about the signaler's competence in the relevant normative domain, information it would be very difficult to garner otherwise. What is conveyed in this case is not only Sarah's competence with respect to a set of behavioral or attitudinal norms, but also her competence in recognizing violations of, as well as the norms for policing, those norms.

- *Commitment*: Signaling does more than just convey the signaler's normative competence. Crucially, it also conveys the signaler's *commitment* to the norm, and this commitment is paradigmatically

expressed in a willingness to enforce the norm.[19] More concretely, Sarah's blame signals that (a) she is disposed to enforce the norm, (b) others are licensed to engage in blame as well; and/or (c) she won't retaliate if others enforce the norm. This information is often intentionally communicated as such, part of the "agent meaning" of our behavioral and attitudinal exchanges (to borrow the language of McKenna 2012: 92-104), although below we will discuss cases of unintentional signaling. A good deal of the complexity of blaming is found in the many delivery devices via which commitment to a norm can be demonstrated. Anger can signal sincere commitment even when words do not, as can certain forms of dispassionate relationship-modification, protest, and communicative demands or invitations. These are the many signaling delivery devices we think the leading theories of blame have (collectively) correctly identified.[20]

- *Self-Disclosure*: In signaling her normative competence and commitments, Sarah also signals an additional range of crucial information—often non-intentionally—about her *agential qualities*. Generally such qualities may be about one's character, judgment, regard for others, values, practical identity, and/or group membership. This is information it would obviously be hard for Tony (and the others) to observe directly about Sarah. However, it is precisely the costliness of the signal—the emotional and enforcement costs (and risked costs) of her blaming him, as well as the more general burdens of being and remaining a blaming agent—that contributes to her signal's reliability to observers. These burdens earn their keep, however. Even if Sarah's blame of Tony, a stranger, in a one-off case like this may generate no benefits for her today, her *commitment* to the set of accepted (and socially approved-of) norms will enhance her reputation in ways that will tend to garner trust, friendship, and beneficial interpersonal transactions in the long run (cf., Frank 1988).[21]

---

[19] A few notes on the "enforcement" clause. First, we use the terms "enforce" and "police" interchangeably. Second, we intend a broad understanding of "enforcement." Sure, sometimes enforcing a norm may involve what may reasonably be construed as sanction or punishment. But many times it doesn't. I don't punish or sanction you when you come home late and drunk yet again and I merely shake my head and shut the door to my room quietly. What I'm nevertheless doing is making it clear that this is a norm I won't tolerate being violated, and the form my intolerance takes aims to ensure both that you are reminded of the importance of the norm and that it will subsequently have a firmer place in your normative deliberations. And this in general is what we will have in mind by "enforcement."

[20] Are there any limits to blame's form and content, though, or is it left utterly arbitrary? If the latter, we would be running roughshod over the important distinctions that blamers themselves take seriously between the wide varieties of disapproving behaviors and attitudes, e.g., indignation, resentment, contempt, disappointment, etc. (We are grateful to an anonymous reviewer for articulating this worry.) In addition to the fact of presumably independent specifications for those behaviors and attitudes, we think there are a variety of limits and conditions on blame's form and content, given human communicative limitations and abilities, restrictions on viable social arrangements, and the presence of other longstanding cultural traditions and norms. There are also numerous features involved in *effective* signaling that will put limits on the contours of blame's form and content, especially if it is to enable its bearers to benefit.

[21] We will say more about the long-term benefits of blame later. For now, though, one might initially worry that this explanation of blame violates some Kantian-style demands that we not use others as mere means to ends they don't share. Given that Tony may not benefit from Sarah's blame but she will, isn't she just using him for an end he doesn't share? (Thanks to an anonymous referee for articulating this worry.) At first pass, no. Sarah is signaling her normative commitments; she's not using Tony for any end, but she is using his wrongdoing as the occasion for marking her normative commitments.

- *Demands and Invitations*: Blame not only signals information; it also often serves to signal imperatives or invitations. Sarah's angry communication to Tony both demands that he recognize he has violated a norm of respect and invites some sort of response from him (e.g., apology, renunciation, etc.).

- *(In)Voluntariness and (Un)Intention*: Sarah's blame signals to Tony are mostly voluntary and intentional: She intends for him to get the information she is conveying. But plenty of blame's signals may be sent regardless of whether they are voluntary or are intended to be what they are. Anger, for example, tends to arise involuntarily, and one may unintentionally send plenty of signals about one's commitment to certain norms (as we will see below) to people who are not the target of one's blame. These non-voluntary, unintentional aspects of signaling generally are quite valuable for tracking cheaters or strategic violators in norm-structured contexts. Because blaming anger and other blame responses are so hard to fake convincingly, they serve as reliable signals of the truthful information that one is really not to be trifled with. It does this in much the same way that one's heated flush signals embarrassment, even if one would prefer to keep it hidden (Frank 1988: Ch. 5; Nichols 2007; McGeer 2013).

- *Multiple Targets, Multiple Signals*: Most accounts of blame have focused on *directed blame*, that is, on dyadic cases where the victim of an offense overtly blames the offender face to face. Even though overt dyadic cases are actually not all that frequent in real life, they still represent to many the prototypical example of blame, by reference to which we are better able to understand non-dyadic examples (McKenna 2013: 121).

  Dyadic cases of directed blame do characterize some of what goes on in our blaming practices, and the signaling feature in such cases is obvious. But we think that that same feature is evident in triadic, quadratic, or larger cases, and this fact is much harder to explain on the leading theories of blame, insofar as a single instance of blame may signal *different things to different audiences*. In content-based theories of blame, the focus is only on the *directed* signal: The blamer angrily protests or communicates something *to* the blamed agent, or blames the offender *to* his or friends (behind the offender's back, say). Other aspects of the multi-dimensional signal are too often elided, ignored, or treated as derivative.

  So notice what happens in our featured case. While Sarah is certainly directing her blame to Tony, she is sending many different signals far and wide. Her signals to Tony are fairly obvious: "I stand up for the norms of respect you just violated, you need to take those norms more seriously, and I'm not to be trifled with." But Sarah is also sending a powerful signal to Bev: "I (Sarah) know what Tony's doing to you is wrong and I've got your back." This is a signal of solidarity (and not a protest, relationship-modification, or communicated demand or invitation *with respect to Bev*). Now suppose that Tony's coworkers are sitting beside him. Surely Sarah's blame of Tony sends a signal to them: "Don't you do this sort of thing either!" And suppose there are bystanders who see the whole exchange. They too are recipients of a signal, one conveying not only information about Sarah's commitments to the norms but also information about what courage looks like. Perhaps her example will empower some of them to do the same in the future (and so what's relevant *for them* aren't reactive attitudes, protest, relationship-modification, or communicated demands/invitations either).[22] Given its

[22] Sometimes we are sufficiently aware of the signaling power of our blame to exploit it, deliberately choosing the audiences for what we know our signals to them will be, strategically coloring the signal, or otherwise being selective about how a signal is likely to be

multichannel nature, in some cases blame's signal might even *exclude the blamed agent altogether*. This is a significant and underappreciated point, for it makes it clear just how distinct blame may be from harsh treatment, sanctions, and punishment of the blamed agent. In such cases of "gossipy" blaming, the blamed agent is oftentimes beside the point. Yet the moral signal can remain crucial for the reputation of the blamer and an important data point for social cooperation. In contrast, this is not so in any instance of harsh treatment, sanctions, and punishment that pretends to moral adequacy.

- *Scope*: Blame's signal is produced by and is a response to only certain kinds of entities. Given the signaling function, in saying that "*x* blames *y*," *x* ranges over all and only those creatures capable of internalizing and being committed to norms, and *y* ranges over all and only those creatures believed by the blamer to be capable of violating norms.[23] So inanimate objects can't blame, and non-agents can't be blamed unless they are (falsely) believed to be the kinds of things that can violate norms. Disagreements about marginal cases—children, for example—are typically disagreements about whether and when the blamed entities are capable of violating norms.[24]

Given the many types of information being signaled, the multiple audiences, and the differing volitional and intentional natures of the wide variety of signals that may be sent, it should be clear why each of the leading theories of blame gets only part of blame's story right. Sometimes their favored mental state or activity produces the requisite signal, but sometimes it doesn't, so none of the theories *alone* captures blame's fundamental feature. The signaling function may be discharged by *any* (or many) of them, depending on the context. Further, the theories also have trouble with false positives, as anger, relationship-modification, protest, or communicative attitudes alone may not be sufficient, in any individual case, for blame. It is typically only when one or more of these mental states and activities have the hard-to-fake costly signaling function we have specified that they will constitute blame. Anger at bad weather doesn't signal one's commitment to violated norms, as the weather doesn't (can't) adhere to norms. Smith's case of the mother who modifies her relationship with her blameworthy son (lowering her expectations for his career) doesn't involve blame unless the mother modifies the relationship in a way that signals her commitment to the norms he violated. Merely protesting someone's behavior (*a la* Sister Helen) isn't blame until it is attached to a signal of one's willingness to *enforce* the relevant norms, and so involves more than merely calling our attention to their violation. And communicating a moral demand or inviting acknowledgment/apology when calling out one's ne'er-do-well

perceived. Think of so-called Facebook Warriors. (See Tosi and Warmke 2016 for a nice discussion of ways in which people signal virtue.) Of course, as we've noted repeatedly, genuine blame is often very hard to fake or manipulate; that's precisely why it has the beneficial effects it has. So strategic and selective signaling is particularly vulnerable to failure. One can often "see through" strategic social media signaling. When this occurs, scorn (or other blaming responses) are likely to arise, which is the cost of dishonesty that makes honest signals reliable (Fraser 2012).

[23] Thanks to David McElhoes for helping us clarify this aspect of the account.

[24] Consequently, when we protest children's actions without blaming them, we are still *teaching* the norms, not yet enforcing them, as the children may not yet have been sufficiently exposed to them to count as violating them. Relatedly, even though we are interested only in the nature of blame, and not its aptness conditions, the constraints on the part of eligible targets of blame we have just noted help explain why some people may be excused or exempted from blame (and so connects up our account to considerations of "moral responsibility"). To *violate* a norm, for instance, may require certain agential capacities and qualities of will that aren't met in cases of accidents, duress, compulsion, mental illness, or, perhaps, psychopathy. Thanks to Robin Zheng and others for discussion.

friends for their wrongdoing isn't blame either until it is connected to a hard-to-fake signal of one's *commitment* to the norms in question (one might laughingly call out one's friends without caring one whit about the norms they have violated).

Several of the above noted features also explain why the functional role of blame is resilient even were one to be aware of its function (see Williams 2002) We ordinarily *want* to be regarded as a normatively competent agent, as a full member of our normative communities. Normative competence, for creatures like us, typically involves seeing the world in certain habitual, norm-structured ways, and the internalization of various norms—including a commitment to their enforcement—is a central part of membership in any such community. Of course, our policing of other norm-violators doesn't always elicit direct contestation with the norm violator. Where social trust is low, the best we may do is to express our frustration to others and invite them to join us in solidarity at our disapproval of the third party. Where social trust is higher, though, it may be that blame can be profitably directed at the offender, and indeed, our reputation for normative competence may depend on direct blame of the offender. In either context, though, discharging blame's function is typically of value to, and valorized by, both the blamer and the blamer's community.[25]

6. Why the signaling theory doesn't dance fancy

The leading theories have difficulty accounting for several important data points about blame. Thus far we have merely presented an alternative account of the unifying core feature of blame. We now need to show how this alternative theory has a much easier job accounting for the four cases motivating the leading theories to dance fancy.

6a. Blaming the Absent

One problem for most other theories is that they think of blame as a species of *holding accountable*. Doing so makes it hard to explain blame of people who are absent, either in different parts of the world or dead. McKenna tries to handle these cases in his conversational theory by explaining how they are intelligible departures from the paradigm case of directed blame. As he puts it, "Blaming in the absence of the blamed can be understood in terms of how the one who blames *would* respond to and . . . converse with . . . the one blamed were the blamer in the presence of the blamed, and were the blamer in a position to alter relevant practices in ways expressive of moral demands, expectations, disappointments, and so on" (McKenna 2012: 177; emphasis ours). We find this to be an implausible way to account for these cases, however.

Suppose you are reading the morning paper and you blame the burglar across town or the mass murderer in Syria you read about, where this involves more than just a belief about their blameworthiness. Typically your blame will occur in the company of someone else. (We'll discuss purely private versions below.) You may call out to your roommate or partner, "You're not going to believe this: That madman in Syria has just used chemical weapons on his own people again!" This is an obvious case of signaling to your roommate or partner: "These are the norms I stand for, and I'd like to hear you say you do too." It's a bonding exercise, as it were.

But were you in the presence of the mass murderer, your blaming response would be *radically* different, we predict, from the mildly irritated shout-out to your roommate or partner in the comfort of your breakfast

[25] Indeed, what may independently provide agent-level justificatory reasons for sticking with the functionally-useful signaling aspect of blame is that participating in it is taken to be reflective of a moral virtue, perhaps something like "uprightness" or "moral courage," or even "moral community member." We are grateful to an anonymous referee for suggesting this line of thought.

nook. This is because what you would be signaling *to the madman* would be quite different were you face to face with him, and how your signaling manifests would also depend on who else was there.

Blaming the dead or historical villains works in a similar way. For instance, how I blame my dead father may take many forms, only some of which are even vaguely similar to how I would actually converse with my father were he alive. What these forms of blame have in common, we think, is that they are just different forms of signaling. I may signal my commitment to various norms my father allegedly violated in many ways, and signal it to many different people, depending on what he did, what it means to me, and the audience to which it may appropriately be expressed.[26] Blaming the dead is mostly a signal for the living.

6b. Dispassionate Blame

Consider a mother who has seen her adult son repeatedly cheat on his romantic partners. She may well blame him without rancor or any other passion: "You keep hurting others and yourself. You exhaust me. I don't know how much more of this I can take." This is an example of blame without reactive attitudes, relationship alteration, or protest (at least of the kind discussed in the literature, as a repudiation of a threatening moral stance). It is indeed part of a communicative exchange. But even so, the communicative theory doesn't capture what is most fundamental to this scenario: This mother is signaling. Not only is she signaling her commitment to (and enforcement of) the norms of relationship-commitment generally, but she may also be signaling *her continuing love and support* for her son (even though she claims it may be on the rocks). Indeed, the latter signal could actually be undermined if she were to signal her commitment to the violated norms with any reactive attitudes, in which case the reactive attitudes version of the communicative theory has trouble accounting for it as well.

6c. Self-Blame

Most standard theories of blame have to dance fancy when it comes to explaining self-blame. After all, it's hard to make sense of how to modify my relationship with myself, or how to protest myself. The reactive attitudes account has a leg up on these others, of course, insofar as self-blame typically involves guilt, but what of cases in which I blame myself without it? I may call myself an idiot for having taken the wrong turn on the way to an appointment, or I might (merely) regret having insulted someone at the bar after having had too much to drink. Perhaps I just give myself a stern talking to in light of some norm violation, reminding myself not to screw up like that in the future.

Is self-blame better described, then, as communicating to myself a demand or invitation for acknowledgment that is rationally intelligible in light of the paradigm of directed blame of others? It seems not. For one thing, I already *know* what's being demanded if I am also the demander (so demanding it would be pointless as a form of communication), and any guilt I feel seems to be the product of my *antecedent* acknowledgment of what I did, and so not an invitation thereto.

In addition, there are plenty of cases in which self-blame seems quite disanalogous from other-blame. One of the most powerful for our purposes is my self-blame for failing to live up to *ideals* that I, and only I, have set for myself—for example, athletic, aesthetic, or religious ideals.[27] But there just is no recognizable notion of directed blame of others for their failures to live up to their *own* such ideals. One might say, of course, that these

[26] We think T.M. Scanlon (2008: 146-7) and Angela Smith (2013: 45) have insightful things to say in this ballpark as well, but we lack space to discuss them here.

[27] McKenna (2012: 73, fn. 14) acknowledges cases like this, although he saves the explanation for why they aren't counterexamples to his thesis for a "later time."

aren't *moral* ideals, so the self-flagellation in response to their violation can't be blame. Now we, the authors, obviously don't think the realm of blame is restricted to the domain of the moral (to be discussed in detail later), but for now let's grant the premise. Suppose instead, then, that I have set very high *moral* ideals for myself, expectations of maximal kindness and generosity that I have failed to live up to today. My (moral) blaming of myself in these cases still has no connection at all to how others would or should respond to me, as no one else has any grounds whatsoever to blame me for failures with respect to my personal ideals, moral or not.

How does the signaling theory improve on these other accounts? Consider first the many cases in which self-blame occurs in front of an audience. Think, for instance, of the quarterback who throws an interception and then pounds his helmet with his fists, the pro golfer who breaks her club in frustration after a poor shot, or the tournament chess player who slumps in agony when he sees how he's just fallen into his opponent's trap. There are clear signals being sent in these cases: "I am committed to norms of excellence that I have failed to live up to." And signaling such things *obviously* has long-term benefits, especially because these kinds of signals are hard to fake, given the kind of serious emotional investment their spontaneous expression typically assumes. People are attracted to those who are deeply committed to excellence (and to self-enforcing it), whether as players on their teams, as objects of rooting, or as their friends.

Signaling a failure of *moral* ideals plays the same role. When I publicly register my disappointment in myself at failing to sacrifice enough of my time for others, I am signaling my virtues: I'm a kind person, and generosity really matters to me. People are more likely to trust and deal with me as a result (as long as I'm not seen as "virtue signaling," though, aiming directly at the benefits by manipulating the signal). These signaling functions have a kind of communicative role, but the cases show just how different blame's signaling function is from conversation or directed communication.

Our way to account for self-blame and its ilk thus far depends on there being an audience to it. But many such cases—as well as cases of other-blame—lack an audience altogether. What are we to say about them?

6d. Private Blame

Private blame—blame that is not outwardly expressed—is an uneasy fit for most theories of blame. But it might also be thought to be a real problem for us, as signals of norm commitment and enforcement would seem to require an audience to generate the cited benefits for the signaler. Alternatively, if we were to deny that there is any such thing as private blame, then our theory would be less about the nature of blame than about *blaming*.[28] Nonetheless, our account has a comparatively straightforward way to accommodate the regularly-occurring phenomenon of private blame: it operates via familiar and predictable mechanisms of norm internalization, which itself a function of signaling.

Let's begin by returning to cases of self-flagellation in which I fail to live up to my own high non-moral competitive ideals, remaining neutral for now on whether this counts as "blame." Suppose that I pound my helmet, or snap my putter in two, or slump in agony over my poor chess move in the tournament. I have clearly internalized some competitive normative ideals in a way that their violation matters to me and so disposes me to respond in these self-enforcing ways, e.g., "Keep your stupid head still when putting!" Now suppose I do these things without any audience, as I may often do. This makes it *particularly* costly. Yet, I am still in all of these cases conveying many kinds of information about my normative commitments—I am sending signals—*even if no one is around to pick up on them*. This, in a nutshell, is the difference between signaling and communicating.

Now revise the scenario to consider my self-flagellating responses to various social ideals that I have *not* set for myself. Beauty standards are illustrative: Societal norms for female beauty are almost impossible not

[28] Our thanks to Angela Smith for raising a form of this concern with us (private communication).

to internalize, as we see media everywhere depicting how women should look, and so those who do not live up to these ideals consequently often feel ashamed. But these ashamed responses will include motivations and activities (e.g., hiding one's face) that, *were* there an audience present, would be recognizable as the ashamed signals they are.

This same basic picture applies to the social norms of *conduct*, which include moral norms. They are determined externally but acquired and internalized fairly early on in childhood development. Once internalized, they operate in various familiar respects from within, disposing one to certain sorts of behaviors, actions, and attitudes. Internalization in many ways simply consists in dispositions to respond in "blamey" ways when the internalized norms are violated. And these are the ways unified by the signaling function we have articulated. If one is truly committed to the norms in question, then one will be disposed to produce a blame signal in response to their violation in all sorts of circumstances, *even when there is no external audience to pick up this signal*.[29] The same is true when one is blocked from expressing those signals, or when one has an overriding interest in suppressing those signals. Even if no one witnesses the torch fishing, or one is hiding one's torch fishing, the act is still what it is *because* of its relationship to the practice that gives it its psychological shape and meaning. Our answer is thus not fancy dancing, as it is for content-based theories of blame. That's because self-blame is an entirely predictable consequence of the genuine norm internalization blame's signaling function puts (evolutionary, cultural) pressure on us to generate. Indeed, this internalization is what generates one of the *costs* of having a genuinely reliable signal, that it causes one emotional turmoil even when no one is around to pick up the commitment signals one is sending.

Just because there is no external audience, though, doesn't always mean there is no received signal. As with the public case, one's private blame may signal to *oneself* where one stands, a reminder of one's virtues or vices, or one's dispositions and normative commitments. Private blame may also enable personal diachronic moral coherence. If I recall that I privately blamed someone for an act I am considering committing, the specter of hypocrisy may moderate the appeal of the action.

An effective signaling system will be rooted in attitudes, subtle facial cues, action tendencies, and bodily movements that are hard to control, which itself makes it hard to deliver false signals and so better enables cheater-detection. But the only way such a system works is if agents are internalizing norms, where that means they are disposed to relatively robust affective and behavioral enforcement responses in a suitably wide range of circumstances. But then internalization is precisely what predictably generates the occurrence of private blame. Again, this isn't an ad hoc epicycle of the account. Rather, it is a straightforward consequence of the psychological demands created by an effective but costly social signaling system, at least in creatures like us.

7. Other Explanatory Benefits of the Signaling Theory

In addition to the four data points the leading theories have to dance fancy around, there are two other data points on the original list that the signaling theory helps to explain very well, we think: hypocritical blame and hypothetical blame. We discuss them both very briefly here. What we say about the second issue will lead into the next section on signaling's domain.

---

[29] Why not view blame simply as the cost of internalizing these norms, a triggered disposition built into one's caring about morality that may be felt or displayed in a wide variety of ways as a *sign*, but not a signal, of one's caring? (This view was suggested to us by Nomy Arpaly and draws from Sher 2006.) The answer is that there are plenty of dispositions that may be activated in virtue of one's caring about morality, not all of which count as blame. As we noted earlier, I may be moved to horror when seeing gross violations of morality, a sign of my caring, but that's not blame.

7a. Hypocritical Blame

There has been a lot written recently about what it means for people to lack the standing to blame others (see, e.g., Smith 2007; Scanlon 2008: 166-79; Duff 2010; Bell 2013). We think our account has a very nice explanation of the main case discussed in this literature: hypocrisy. Suppose Aara cheats on her husband Donald, and after he finds out he cheats on her too, and then Aara blames him for doing so. Something is "off" about her blaming him (Donald might be moved simply to sputter or laugh in response: "Who are you to blame me?"). But what is off about it? After all, Donald did violate the norms of their relationship, and so Aara's anger would be fitting, she would have grounds for modifying her relationship with him in light of what he did, and both a protest and her communication of the moral demands in play would be appropriate. None of these responses' aptness would be undermined by her hypocrisy.

What *would* be undermined is the integrity of her blame's signal. Quite simply, her hypocrisy casts doubt that the signal she is sending via blame conveys accurate information about her commitment to the (enforcement of the) norms. So she could puff up her chest and angrily shout at Donald, demanding repudiation and acknowledgment till the cows come home. In doing so, she would be engaged in genuine blame. Given her recent unapologetic infidelity, however, Donald still has significant reason to see her commitment to the norms of fidelity, trust, and loyalty as suspect, and thus, her blame (especially in this case) as another deceptive signal. Insofar as the signals hypocritical blame sends are unreliable (or are aptly perceived as such), there is, in a sense, no *point* to sending them. What is "off" about hypocritical blame is less about the standing to blame and more about the nature and integrity of its signal.

7b. Hypothetical Blame

A special advantage of our approach is that it elegantly explains cases of what we may think of as blame where *none* of the theories has obvious answers. Consider the example from *Force Majeure*, in which one character seems to blame her boyfriend for what she predicts he *would* do (abandon his family to scramble for his own survival), not for anything he did. In this case, he did not violate any norm. But there is nevertheless signaling: His girlfriend is conveying information about her commitment to, and, crucially, her willingness to enforce, certain norms of their relationship she sees as possibly *threatened*, perhaps given his character. She may also be signaling her ideals about not being a chump.

Now we admit that this case may not clearly count as blame for some people. Reasonable people may disagree. Our contention is that once we appreciate the wider features of blame, then whether or not this is a case of blame (as it is presently understood) is mostly beside the point. We make the case for this point in the next section.

8. Moral and Nonmoral "Blame"

It is very often claimed that blame is a response to wrongdoing, where what people have in mind is *moral* wrongdoing (Wallace 1994; Watson 1987; Hieronymi 2004; Nelkin 2011; McKenna 2012; and many more). The way it is typically put suggests that moral wrongdoing is a necessary condition of blame. Indeed, it is the presumed connection to moral wrongdoing that is sometimes thought to lend blame a distinctive force and feel, insofar as being blamed for moral violations cuts closer to the bone than criticism for other sorts of normative violations (see, e.g., Wallace 1994; Scanlon 1998: Ch. 6; Hieronymi 2004; Smith 2008; Cogley 2013; Bennett 2013).

We think this is a mistake, belied by compelling examples. Suppose in the heat of a close game, a soccer coach angrily pulls a defensive midfielder who allowed a striker to get by her and score. The coach berates his player, yelling at her up and down the sidelines for a few minutes. Is his response *blame*? It seems obvious that it

is, even though it is quite implausible to insist that she violated any *moral* norms. We can even stipulate if you like that the player loves and respects her coach, it is the final game of the season, and it is blowout. The coach has nevertheless warned his underperforming player that she'll be benched if she keeps letting offensive players past her. When she lets through a second goal, the coach berates her as above. What moral norms could she have violated? In truth, she just wasn't up to snuff athletically today. Still, every other familiar feature of ordinary blame is in place: The coach is angrily communicating to his player, protesting her poor play, and modifying their relationship. And what unites all of these features? Signaling, and lots of it: He is signaling to the blamed player, to her teammates, and to the audience his high standards and his willingness to enforce these norms by making an example of players who fail to live up to them.

This sort of activity happens all of the time, and in many (all?) nonmoral normative domains (see also Williams 1995: 40, with whom we agree). One screenwriter may criticize another for failing to follow through on the political statement implicit in early scenes of his movie script, signaling her own commitment to demanding aesthetic norms. Or a high society matron may with a flash of her eyes scold another who has used her dessert fork for the main course, signaling her commitment to certain norms of etiquette. And there are many televised examples of chefs berating their cooks for not making high enough quality dishes. Whatever the domain, each is an example of someone signaling a commitment to, and a willingness to enforce, a norm. The core phenomenon is precisely the same. Insisting that blame is exclusively moral obscures this fact.

Perhaps, though, some will want to insist that such non-moral cases aren't examples of REAL blame. Either of two responses are available. First, the bold response: *Why does it matter?* Whether we call these things "blame" or merely "blamey," the signaling aspect unifying all uncontroversial cases of moral blame is also found in our responses to norm violations in plenty of non-moral normative domains. Indeed, what the hypothetical "blame" case in the previous section suggests is that signaling of our sort may even be present in cases *without* any norm violations yet, where responses serve primarily to serve notice that "I've got my eye on you." That is to say, the policing feature of the signaling function is not merely retroactive; it is sometimes *pro*active, in the way that good actual policing should be. Likely the best way to enforce a norm is to ward off its being violated in the first place.[30] By these lights, it is unclear what hangs on regimenting blame's domain in the suggested way.

A second and more conciliatory response can allow that, as a matter of current usage, there may be some conceptual or natural language constraints on blame. Perhaps it is simply a conceptual or linguistic truth in our community that blame requires that there be a norm violation, and perhaps even a *moral* norm violation, for something to count as blame. Even so, it is the functional structure we have identified, namely, the signaling role, that does the explanatory heavy lifting. Given that those signaling structures clearly exist across moral and non-moral contexts, it is an insight delivered by our theory that moral blame is functionally indistinguishable from signaling behaviors we perform in nonmoral contexts. Here, natural language is revealed as a poor guide to the deep structure of what is going on when we blame.

We are of two minds about which of these responses is the more promising. Both foreground the explanatory power of the functional account. The bold response has a clear theoretical elegance to it. The conciliatory response can accommodate some conceptual and linguistic constraints on blame, but at the cost of requiring a further story about the relationship between the conceptual requirements and the signaling function. Whatever the right answer is here, we are of one mind in insisting that moral blame is clearly of a piece with the structure of what is most naturally called non-moral blame. What matters most for our purposes is simply that we bring to the fore the signaling function that is the crucial core of the paradigm forms of both, a function that

[30] Macnamara 2013 (160-61) also presents an interesting case along these lines.

has for too long been overlooked. Signaling of this sort is found in a wide array of social interactions, moral and non-moral, and may well be a crucial sort of interpersonal glue. Whether its every instantiation has to be labeled "blame," though, is just not our concern. To fight about the term is to miss the signal.

9. Coda: Instructions on how to douse the torch's flame

We believe the signaling theory of blame does at least as well as any going competitor account, and that this fact should be enough for us to treat it as a contender approach. Naturally, we think it does better than the average competitor account, but disagreement is to be expected. Still, in the spirit of constructive engagement we conclude by pointing out two ways one could offer philosophically interesting challenges to the signaling account.

First, one might advance an alternative content-based theory of blame's crucial core, an account avoiding the fancy dancing we think is endemic to such views. This would have to include provision of clear cases of blame without the relevant connection to signaling, an explanation of why the seemingly necessary connection to signaling in paradigm cases of blame does no heavy explanatory lifting, and how the relevant content-based view accounts for the blame data set at the beginning without the usual fancy dancing. We are unsure what such a view might look like, and we are inclined to think that even conciliatory versions of the signaling view that allow for some conceptual constraints on blame have work to do in motivating and justifying ongoing commitment to that content.[31]

A second way to challenge the account would be to agree with us that blame has a functional core but to disagree with us on what its function is. For example, suppose something like McKenna's conversational account (2012), Smith's protest account (2013), or Fricker's communicative account (2014), were best construed as offering broadly functional accounts of blame. If so, then they would be competitors to the present proposal and they would have to be assessed in the usual fashion: We would highlight the comparative advantages of our account, the competitor functionalist theorists would highlight the virtues of their accounts, and then we would determine which account has a greater net balance of virtues over vices. We think we have certainly provided sufficient reasons to prefer our functional theory to a content-based conversational, communicative, or protesting theory in this paper, but we haven't yet wrestled explicitly with the alternative *functional* versions of the theories, and we allow that there are likely other possibilities we haven't surveyed at all. But we certainly look forward to having those conversations.[32]

---

[31] Gunnar Björnsson has floated one such possibility to us in conversation that we would like to see fleshed out some more. The basic idea is to detach the enforcement aspect from our proposed signaling function and instead make *it* the core feature of a content-based account. Blame, then, would be whatever mental state or activity (paradigmatically?) enforces norm violations. This is a promising start, but hard questions remain. For instance, why think that this view won't have to engage in fancy dancing to deal with the standard cases of private blame, or blaming the absent or dead? And what of cases of enforcement of the norms without any commitment to them (as when one does so for purely prudential reasons)? Isn't enforcement without investment possible? At any rate, our aim has been merely to start a conversation like this, not finish it, and we are grateful to Björnsson for starting it in these terms.

[32] The authors are grateful to many audiences to which earlier versions of this material were presented, including those at Gothenburg University, UNC-Greensboro, the 2017 Pacific APA, the 2016 meeting of the American Association of Mexican Philosophers, the Southern California Agency and Responsibility Workshop, the Moral Psychology Research Group, Arizona State, and Hebrew University. We are especially grateful to our commentators at the APA invited symposium: Nomy Arpaly, Victoria McGeer, and Robin Zheng. For discussions and comments on earlier drafts, we are grateful to David Brink, Teresa Bruno, David McElhoes, Robert Hartman, Michael McKenna, Dana Nelkin, Shaun Nichols, Douglas Portmore, Angela Smith, Dan Speak, as well as a particularly helpful anonymous referee. The idea for this paper came about during discussions we started in Gothenburg, Sweden, during an early conference at the Gothenburg (now Lund) Responsibility Project, and we are grateful to the organizers, Paul Russell

and Gunnar Björnsson, for that opportunity. We are also grateful to the University of San Francisco for providing funding to enable us to work in person together for a few days over a couple of years when we were writing the paper in 2015-2016.

REFERENCES

Arpaly, Nomy. 2006. *Merit, Meaning, and Human Bondage*. Princeton, NJ: Princeton University Press.

Arpaly, Nomy and Schroeder, Timothy. 2014. *In Praise of Desire*. Oxford: Oxford University Press

Axelrod, Robert. 1986. "An Evolutionary Approach to Norms." *The American Political Science Review* 80: 1095-1111.

Bell, Macalaster. 2013. "The Standing to Blame: A Critique." In Coates and Tognazzini 2013b, pp. 263-281.

Bennett, Christopher. 2002. "The Varieties of Retributive Experience." *The Philosophical Quarterly* 52 (207): 145–63.

——. 2013. "The Expressive Function of Blame." In Coates and Tognazzini 2013b, pp. 66-83.

Bird, Rebecca Bliege, Smith, Eric, and Bird, Douglas W. 2001. "The hunting handicap: costly signaling in human foraging strategies." *Behavioral Ecology and Sociobiology* 50: 9-19.

Bird, Rebecca Bliege, and Smith, Eric. 2005. "Signaling Theory, Strategic Interaction, and Symbolic Capital." *Current Anthropology* 46: 221-248.

Calhoun, Cheshire. 1989. "Responsibility and Reproach." *Ethics* 99: 389-406.

Carlsson, Andreas Brekke. 2017. "Blameworthiness as Deserved Guilt." *The Journal of Ethics* 21: 89-115.

Clarke, Randolph, McKenna, Michael, and Smith, Angela M., eds. 2015. *The Nature of Moral Responsibility*. New York: Oxford University Press.

Coates, D. Justin, and Tognazzini, Neal. 2013a. "The Contours of Blame." In Coates and Tognazzini 2013b, pp. 3-26.

Coates, D. Justin, and Tognazzini, Neal, eds. 2013b. *Blame: Its Nature and Norms*. Oxford: Oxford University Press.

Couch, Mark B. 2017. "Causal Role Theories of Functional Explanation." *The Internet Encyclopedia of Philosophy*. URL: http://www.iep.utm.edu/func-exp/

Cummins, R. 1983. *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

D'Arms, Justin, and Jacobson, Daniel. 2000. "The Moralistic Fallacy: On the 'Appropriateness' of Emotions." *Philosophical and Phenomenological Research* 61: 65-90.

Darwall, Stephen. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

Driver, Julia. 2015. "Appraisability, Attributability, and Moral Agency." In Randolph Clarke, Michael McKenna, and Angela Smith, eds., *The Nature of Moral Responsibility* (New York: Oxford University Press), pp. 157-174.

Egas, M and Riedl, A. 2008. "The Economics of Altruistic Punishment and the Maintenance of Cooperation." *Proceedings of the Royal Society B*. 275: 871-878.

Fehr, E. and Gächter, S. 2002. "Altruistic Punishment in Humans." *Nature* 415.10: 137-140.

Fine, Cordelia, and Kennett, Jeanette. 2004. "Mental Impairment, Moral Understanding and Criminal Responsibility: Psychopathy and the Purposes of Punishment." *International Journal of Law and Psychiatry* 27: 425-443.

Fowler, J. 2005. Altruistic punishment and the origin of cooperation." *Proceedings of the National Academy of Sciences of the United States* 102.19: 7047-7049.

Frank, Robert H. 1988. *Passions with Reason*. New York: W.W. Norton & Co.

Franklin, Christopher. 2013. "Valuing Blame." In Coates and Tognazzini 2013, pp. 207-223.

Fraser, Ben. 2012. "Costly signalling theories: beyond the handicap principle." *Biology & Philosophy* 27: 263-278.

Fricker, Miranda. 2014. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs*.

Frijda, Nico. 1986. *The Emotions*. Cambridge: Cambridge University Press.
——. 2007. *The Laws of Emotion*. Mahwah, NJ: Erlbaum.
Gintis, H. 2003. "The Hitchhiker's Guide to Altruism: Gene-Culture Coevolution, and the Internalization of Norms." *Journal of Theoretical Biology* 220, 407-418.
Glover, Jonathan. 2014. *Alien Landscapes?: Interpreting Disordered Minds*. Cambridge, MA: The Belknap Press of Harvard University Press.
Hieronymi, Pamela. 2001. "Articulating an Uncompromising Forgiveness." *Philosophy and Phenomenological Research* 62: 529-555.
——. 2004. "The Force and Fairness of Blame." *Philosophical Perspectives* 18: 115-148.
Kelly, Daniel, and Erica Roedder. 2008. "Racial Cognition and the Ethics of Implicit Bias." *Philosophy Compass* 3, no. 3: 522–40.
Koleva, Spassena P., et al. 2012. "Tracing the Threads: How Five Moral Concerns (Especially Purity) Help Explain Culture War Attitudes." *Journal of Research in Personality* 46: 184-194.
Levy, Neil. 2007. "The Responsibility of the Psychopath Revisited." *Philosophy, Psychiatry and Psychology* 14: 129-138.
Macnamara, Coleen. 2013. "Taking Demands Out of Blame." In Coates and Tognazzini 2013b, pp. 141-161.
——. 2015. "Blame, Communication, and Morally Responsible Agency." In Clarke, McKenna, and Smith 2015, pp. 211-235.
Mauss, Marcel. 1924. *The Gift: Forms and Functions of Exchange in Archaic Societies*. London: Cohen and West.
Maynard-Smith, J. and Harper, D. 2003. *Animal Signals*. Oxford: Oxford University Press.
McGeer, Victoria. 2013. "Civilizing Blame." In Coates and Tognazzini 2013b, pp. 162-188.
McKenna, Michael. 2012. *Conversation and Responsibility*. New York: Oxford University Press.
——. 2013. "Directed Blame and Conversation." In Coates and Tognazzini 2013b, pp. 119-140.
——. Forthcoming. "Power, Social Inequities, and the Conversational Theory of Moral Responsibility." In Katrina Hutchison, Catriona Mackenzie, and Marina Oshana, eds., *Social Dimensions of Moral Responsibility*.
McKenna, Michael and Vadakin, Aron. Review of George Sher's *In Praise of Blame*. *Ethics* 118 Pp. 751-756.
Nelkin, Dana. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.
——. 2015. "Psychopaths, Incorrigible Racists, and the Faces of Responsibility." *Ethics* 125: 357-390.
Nichols, Shaun. 2007. "After Compatibilism: A Naturalistic Defense of the Reactive Attitudes." *Philosophical Perspectives* 21: 405-28.
Nussbaum, Martha C. 2016. *Anger and Forgiveness: Resentment, Generosity, Justice*. Oxford: Oxford University Press.
Portmore, Douglas W. Forthcoming. "Control, Attitudes, and Accountability." *Oxford Studies in Agency and Responsibility* 6.
Scanlon, T.M. 2008. *Moral Dimensions*. Cambridge, MA: Belknap Press of Harvard University Press.
Scarantino, Andrea. 2014. "The Motivational Theory of Emotions." In Justin D'Arms and Daniel Jacobson, eds., *Moral Psychology and Human Agency*. Oxford: Oxford University Press, pp. 156-185.
Searcy, W, and Nowicki, S. 2005. *The Evolution of Animal Communication*. Princeton: Princeton University Press.
Sher, George. 2006. *In Praise of Blame*. Oxford: Oxford University Press.
Shoemaker, David. 2007. "Moral Address, Moral Responsibility, and the Boundaries of the Moral Community." *Ethics* 118: 70-108.

——. 2013. "Blame and Punishment." In Coates and Tognazzini 2013, pp. 100-118.
——. 2015. *Responsibility from the Margins*. Oxford: Oxford University Press.
——. 2018. "You Oughta Know! Defending Angry Blame." In Myisha Cherry and Owen Flanagan, eds., *The Moral Psychology of Anger*. London: Rowman & Littlefield. pp. 67-88.
Smith, Angela M. 2007. "Being Responsible and Holding Responsible." *The Journal of Ethics* 11: 465-484.
——. 2013. "Moral Blame and Moral Protest." In Coates and Tognazzini 2013b, pp. 27-48.
Sosis, Richard. 2001. "Costly Signaling and Torch Fishing on Ifaluk Atoll." *Evolution and Human Behavior* 21: 223-244.
Spence, M. 1973. "Job market signaling." *Quarterly Journal of Economics,* 87: 355-374.
Spence, M. 2002. "Signaling in retrospect and the informational structure of markets." *American Economic Review* 92: 434-459.
Strawson, P.F. 2003. "Freedom and Resentment." In Gary Watson, ed., *Free Will*, Second Edition (Oxford: Oxford University Press), pp. 72-93.
Talbert, Matthew. 2012. "Moral Competence, Moral Blame, and Protest." *The Journal of Ethics* 16: 89-109.
Thrasher, John, and Handfield, Toby. 2018. "Honor and Violence: An Account of Feuds, Duels, and Honor Killings." *Human Nature* 29: 371-389.
Tosi, Justin, and Warmke, Brandon. 2016. "Moral Grandstanding." *Philosophy & Public Affairs*. DOI: 10.1111/papa.12075
Vargas, Manuel. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.
Veblen, Thorstein. 1994/1899. *The Theory of the Leisure Class*. New York: Dover.
Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
——. 2011. "Dispassionate Opprobrium: On Blame and the Reactive Sentiments." In Wallace, Kumar, and Freeman 2011, pp. 348-372.
Wallace, R. Jay, Kumar, Rahul, and Freeman, Samuel. 2011. *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*. Oxford: Oxford University Press.
Watson, Gary. 1987. "Responsibility and the Limits of Evil." In *Responsibility, Character, and the Emotions*, edited by Ferdinand David Schoeman, New York: Cambridge, pp. 256–86.
——. 2004. *Agency and Answerability*. Oxford: Oxford University Press.
——. 2011. "The Trouble with Psychopaths." In Wallace, Kumar, and Freeman 2011, pp. 307-331.
Williams, Bernard. 2002. *Truth and Truthfulness: An Essay in Genealogy*. Princeton: Princeton University Press.
Wolf, Susan. 2011. "Blame, Italian Style." In Wallace, Kumar, and Freeman 2011, pp. 332-347.
Zahavi, Amotz and Zahavi, Avishag. 1997. *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.