


ARTICLE

# The Epistemology of Moral Praise and Moral Criticism

Jimmy Alfonso Licon 

The Mercatus Center at George Mason University, Arlington, Virginia, USA  
Email: [jimmylicon01@gmail.com](mailto:jimmylicon01@gmail.com)

(Received 11 November 2020; revised 27 May 2021; accepted 23 June 2021;  
first published online 9 September 2021)

## Abstract

Are strangers sincere in their moral praise and criticism? Here we apply signaling theory to argue *ceteris paribus* moral criticism is more likely sincere than praise; the former tends to be a higher-fidelity signal (in Western societies). To offer an example: emotions are often self-validating as a signal because they're hard to fake. This epistemic insight matters: moral praise and criticism influence moral reputations, and affect whether others will cooperate with us. Though much of this applies to *generic* praise and criticism too, moral philosophers should value sincere moral praise and moral criticism for several reasons: it (i) offers insight into how others actually view us as moral agents; (ii) offers feedback to help us improve our moral characters; and (iii) encourages some behaviors, and discourages others. And so as *moral agents*, we should care whether moral praise and moral criticism is sincere.

**Keywords:** Signaling theory; moral criticism; moral praise; cooperation

## 1. Introduction

Are strangers sincere in their moral praise and criticism? This question *morally* matters: moral praise and criticism influence our moral reputations, which in turn influence who will cooperate with us. Since long-term cooperation is *practically* valuable, allowing us to do things we couldn't otherwise, third-parties will likely assume apparently sincere moral evaluations are at least a decent indication of one's moral reputation than insincere ones. Humans depend on each other, facilitated by moral reputation; preserving our moral reputation is highly valuable. Further, sincere moral praise and criticism can be *morally* valuable too: it can do things like provide genuine moral feedback of what others think of us – sometimes strangers can see things about us that those closest to us simply cannot – which allow us to improve as moral agents. And yet despite the high stakes, it can be hard to tell if strangers are sincere in their praise and criticism.

This paper defends an answer, to the question we posed, using signaling theory: some signals are cheap, others are costly, and the costliness of a signal indicates the fidelity of that signal. A fruitful approach to the question of the sincerity of moral praise and criticism is to treat them as signals with varying fidelity. Though our thesis needs clarification to distinguish between two different kinds of criticism – which we offer in a later section – we can state it as follows:

*Costlier*

Moral criticism is often a costlier signal than moral praise partly because it risks retaliation; if we know nothing else about the person moral criticizing and praising, then *ceteris paribus*, we should treat moral criticism as more likely to be sincere than moral praise.

We should think of moral praise as a *cheaper* signal, as insincere moral praise is less costly, e.g., complimenting the food as a dinner guest to be polite isn't particularly risky. People have many reasons to morally praise others unrelated to whether they acted in a morally praiseworthy manner – it's nice to be told one is nice. However, even with good reasons to morally criticize others, we often don't criticize to avoid blowback. We should rather expect individuals to risk retaliation by morally criticizing others in rare cases; e.g. when they believe someone deserves it (and maybe not even then). And unlike with moral praise, people who don't criticize immoral behavior can be seen as second-order *moral* free riders (Yamagishi 1986).

We must bear in mind that *Costlier* is an epistemic heuristic that applies chiefly to moral praise and moral criticism from strangers; where we know little about their intent other than the costliness of the respective moral signals. Moral criticism is more likely to be a high-fidelity signal than moral praise in light of the salient incentive structure – moral criticism particularly can seriously damage one's moral reputation. However the same cannot be said of moral praise: subjects of moral praise largely lack the incentive to retaliate. Third-parties may sometimes retaliate against someone for their moral praise, but this applies to moral criticism too. And since we should treat moral criticism and moral praise as arguably equally interpersonally, but not equally personally, costly, moral criticism has higher costs overall.

Although economists pioneered incentive-based epistemology (Cowen 2012), philosophers have followed suit (Moller 2013; Kogelmann and Wallace 2018). This paper adds to that growing literature. However, insights from economics don't just stop at epistemology: in a recent paper, Shoemaker and Vargas (2019) defend a signaling theory of *blame*. The idea here is that blame – in reaction to norm violations – should be framed in terms of function: the function of blame is to provide a costly signal that the blamer is committed to the violated norm, and to the punishment of those who violated that norm. It is costly since it risks retaliation by the target of the blame. By example: by blaming someone for violating a norm, Samantha is signaling that she is willing to pay a price to see the norm violator punished, namely by risking retaliation for blaming the norm violator, and thus signals her commitment to the norm. So in that respect, the signaling theory of blame from Shoemaker and Vargas shares some commonality with the thesis of this paper, namely: costly signals aid in regulating our moral lives.

We should clarify the scope of our thesis. Our thesis relies on studies that use WEIRD subjects (Western, Educated, Industrialized, Rich, and Democratic; see Henrich 2010), and so we should be careful *not* to generalize from such studies to human psychology broadly. We don't know if they apply to humans writ large, or just to humans who are WEIRD. This is why *Costlier* is pitched as a *heuristic* sensitive to cultural differences: it is arguably a reliable heuristic, but only so far as we know when broadly applied to individuals in the West. It isn't much of a criticism to point out that there are non-Western cultures where *Costlier* would be a poor heuristic, as this would be to overextend *Costlier* beyond its intended scope.

We'll proceed as follows. First, we briefly discuss signaling theory, and then use some examples to illustrate how it works. Second, I apply signaling theory to moral praise and moral criticism, and then argue moral criticism is often (though not always) a high-fidelity signal – one that we can be assured is often enough sincere – moral praise

tends to be a lower-fidelity signal, and that the fidelity of each signal is indicative of the sincerity of moral praise and moral criticism. The underlying incentives, and degree of costliness of moral criticism and moral praise as signals, can help determine the sincerity of that praise and criticism.

## 2. The nature of signaling

Signaling theory, discovered independently in evolutionary biology and economics, explains behavior and biological features as a signal of qualities that aren't otherwise obvious (Spence 1973; Zahavi 1975). Peacocks are a classic case of signaling: their colorful plumage would handicap avoiding predators if they were unfit. Engagement rings are another example: genuine ones are expensive, and jewelers hard to fool. They are a high-fidelity signal of a sincere intent to marry. Costly signals highly correlate with the quality they represent: seeing such a signal can assure us that the quality is present. Other signals are less reliable; they can be faked by those without the quality, but who want to fake it. A stable signaling system has the following features:

- (1) Some members of a group have a quality that is hard to perceive directly, but to which a reliable signal could attach.
- (2) There are observers who would benefit 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 observer's expense.
- (4) The cost of the signal must have some benefit to the signaler (Bird and Smith 2005: 224).

We should recognize the aim of a signaling system is to provide a *high-fidelity* signal to indicate qualities that would otherwise be missed. High-fidelity signals come in many forms: costly, self-verifying, and involuntary. *Costly signals* are hard to fake. One popular way to think about costly signals is in terms of handicapping: something that makes it hard to send a signal without the quality in question. *Self-verifying signals* are verified by their presence. The ability to lift two-hundred pounds over one's head is a self-verifying signal of strength: one cannot lift that much weight without the requisite strength. *Involuntary signals* cannot be easily faked; emotions, like anger and relief, are good examples (Frank 1988). Kay's moral outrage is a high-fidelity signal that she believes something is wrong in that outrage is hard to fake.

We already have background evidence that *one* of the functions of morality – though hardly the only function of morality – is to signal moral qualities to others. This signaling component of morality makes sense in light of the fact that humans are social creatures who depend on cooperation from others to survive. And thus how we *look to others* matters; to be deprived of cooperation from others is on par with death. Many ancients saw banishment from society as worse than death: death is inevitable here too, but prolonged. It shouldn't be shocking then that people often prefer 'jail time, amputation of limbs, and death to various forms of reputation damage' like acquiring a reputation as a Nazi or child molester (Vonasch *et al.* 2018: 604; see also Sperber and Baumard 2012), or kids as young as five prefer to maintain a good reputation, and 'refrain from cheating at the cost of losing a highly desirable prize' (Fu *et al.* 2016: 277).

We already know cooperation is modulated by reputation (Henrich 2015), moral reputations are a major factor in mate selection (Miller 2007), and signaling is pervasive in cooperation and human behavior generally (Simler and Hanson 2018). It thus

squares with what we already know that moral criticism and praise would have a social function too; they would modulate how third-parties see us. Our project applies aspects of signaling theory to moral praise and moral criticism. Moral praise and criticism signal, to various degrees, how one's moral reputation is viewed by others, and such signals have an asymmetric aspect: moral criticism is often a higher-fidelity signal, and thus more likely sincere, than moral praise. We turn to that next.

### 3. Moral praise and moral criticism

The value of moral praise and moral criticism as signals should be clear given the cooperative nature of humans. The value of cooperation helps to explain why we care so much about the moral reputations of our cooperative partners – especially if cooperation is long-term like, say, choosing a spouse or business partner. Moral praise and criticism are reputational inputs; and so, the sincerity of the praise and moral criticism matters. If moral criticism is often a higher-fidelity signal than moral praise, we can weigh those moral signals suitably. Just as the peacock can't signal fitness with colorful plumage without risking predation too, it is hard to morally criticize someone without risking their wrath.

#### 3.1. *The source of costliness*

If we frame moral praise and moral criticism as cheap or costly signals, we should understand what makes them costly or cheap. We should first appreciate why moral criticism can be costly to clarify why moral praise isn't likely to be as costly as moral criticism. Here are some reasons to think it would risk retaliation from the subject of that criticism:

##### *Self-conception*

People tend to self-represent as good. Moral criticism can result in psychological discontent by threatening to undermine the coherence of that self-conception (Hardy and Carlo 2011).

##### *Sanction and Punishment*

Moral criticism increases the odds that the subject of criticism will be punished by others, and thus motivates retaliation; sanctions and punishment, and mechanisms like third-party punishment, meta-punishment, aid in norm enforcement (Boyd *et al.* 2003; Jordan *et al.* 2016).

##### *Lack of cooperation*

Moral criticism can indicate to third-parties that the subject of moral criticism would be a poor cooperative partner; people typically prefer to cooperate with individuals who are subject to less moral criticism, *ceteris paribus* (Miller 2007).

We are now in a position to make an important distinction between two different *types* of moral criticism – following an influential paper by Watson (1996): criticism of actions (accountability criticism), and criticism of attitudes and character (attributability criticism). The former kind of criticism has to do with the actions one performs: we would criticize Omar, say, since he stole someone's wallet, and perhaps demand he be punished; this is a case of accountability criticism. Whereas, the latter kind of criticism targets the attitudes and character of someone: we may think Sam deserves criticism for his reactionary political attitudes, even if we don't think that he should be

punished for them – though perhaps he should be punished if he *acted on them*. This would be a case of attributability criticism.

With this distinction in hand, we can revisit the source of costliness question with respect to different kinds of criticism. The sources of costliness will vary somewhat depending on whether we are talking about criticism of actions, or criticism of attitudes and character. Start with the first source of costliness: the target of criticism wanting to preserve their self-conception as a good person. If someone criticized Bob for acting badly while in a bad mood – perhaps he was rude to a waiter because of hunger and exhaustion – it isn't clear one would risk retaliation here beyond a nasty retort by Bob since one-off criticism of a rude comment may not threaten Bob's self-conception as a good person. Of course, if Bob was routinely rude to those around him, as a regular feature of his personality, than strangers regularly and independently pointing this out may threaten his self-conception as a good person. In contrast, suppose someone criticized Bob for his cruel political beliefs or lack of honesty; criticism of this sort may challenge Bob's self-conception as a good person, risking retaliation by Bob.

Now consider that sanctions and punishments are often (but not always) reserved for actions. We should expect that accountability criticism – where one is criticized for their deeds – would be more likely to encourage punishment from others, and thus risk retaliation. We should thus expect that the possibility would be a source of costliness, by encouraging retaliation, when the criticism is directed at an action, and thus may encourage punishment by others. By example, we wouldn't expect Republicans to punish Bob because he strongly supports a woman's right to choose. So while there is a risk the target of attributability criticism will be sanctioned for their beliefs and character, we should primarily expect that the source of costliness, with respect to punishment, lies mostly with accountability criticism (based on what someone *does*).

Finally, criticizing someone – either for their actions or beliefs and character – can cripple their access to the cooperation market. Of course there is nuance operating here as there often is with complex social rules and interactions, but by and large, we would expect that when criticizing someone for their actions, or their beliefs and character, they thus risk losing out on potential cooperators, sexual partners, friends, and whatnot. This will, of course, depend on the details: perhaps Bob isn't shut out of the cooperation market because of a rude comment he made while hungry and tired; but we can see how regularly treating others rudely may harm Bob's chances of securing cooperation partners. And the same holds for beliefs and character: depending on the details, folks may not want to cooperate with Bob if he has cruel political or moral beliefs, say, because of how others may view it. So with respect to the sources of costliness, we should expect accountability-criticism to be costlier than attributability-criticism, *ceteris paribus*.

Additionally, we have indirect evidence that supports *Costlier* relating to the conveying of bad news. There is good empirical evidence of what psychologists call the 'MUM effect': people are often reluctant to transmit bad news – a message communicating information anticipated by the transmitter to be negatively valenced and unknown by the recipient – to others for fear that it will have bad consequences for them (Rosen and Tesser 1970; Dibble and Levine 2013). The psychological factors here resemble those of moral criticism. As researchers who investigated the 'MUM effect' and its relationship to rumor transmission write:

Overall, it can be concluded that the definitiveness of the consequences of the [bad] news and the relationship between communicator and recipient increase the likelihood of bad news transmission, possibly because they increase the perceived moral responsibility to transmit the news. Both factors *do not affect the likelihood of good news transmission*. (Weenig *et al.* 2001: 460; my emphasis)

Although people tend to pass along rumors, which are mostly negative, the empirical evidence shows that this is because rumors are often transmitted amongst people in a social circle. However, people are increasingly reluctant to transmit bad news as the recipient of the bad news is less and less known to them; they would thus be the least likely to transmit bad news to strangers, *ceteris paribus*. The dynamics are similar to moral criticism: especially when one is sincere, moral criticism can function as bad news – it can reveal one sincerely thinks that you’ve done something bad or have bad moral character, and damage your moral reputation. Of course, it should need not be said that bad news doesn’t have an *evaluative* character to it like moral criticism does, but it still illustrates something important on the sender side of bad news, whether conveying generic bad news or moral criticism: if people don’t like the message you’re sending, they may have incentive to retaliate.

Further, moral criticism, to a larger degree than praise, makes the subject *accountable* to others. Thus it seems moral criticism can be costly, by inviting retaliation against the target of that criticism – whereas, the reasons above don’t apply to *the subject of the moral praise* to retaliate against the source of that praise. We would often like to be morally praised; it may improve our moral reputations. Even if moral praise fails to improve one’s moral reputation – e.g., everyone else knows Molly is a terrible person, moral praise from a passer-by aside – it is unclear how moral praise could harm one’s moral reputation; it is either positive or neutral for the subject.

Moral praise can be costly sometimes too. Just by example, Nelly lost her job at the *New York Times* by praising the Armenian genocide in her editorial. There are cases where moral praise can be costly, but that costliness is often (but not always) from third-parties: moral praise and moral criticism can be *interpersonally* costly, while moral criticism tends to be *personally* costlier than moral praise. We turn to that issue next.

### 3.2. Varieties of costliness

We should distinguish different sources of costliness. When Sally morally criticizes Bob’s meat eating, for instance, she faces potential backlash from a couple of sources. The first is from Bob: he may have reason to retaliate against Sally – e.g., it threatens Bob’s self-conception as a good person. Second, there is potential third-party retaliation: friends, peers, and others friendly to Bob (or meat eating) may retaliate against Sally; her criticism of meat eating may threaten the livelihood of Bob and his coworkers. We should, when assessing *Costlier*, distinguish between the *personal* and *interpersonal* costliness of moral praise and moral criticism:

**Personal Costliness:** The risk the subject of moral praise and criticism will retaliate against the person who praises and criticizes them.

**Interpersonal Costliness:** The risk that third-parties, distinct from the subject of moral praise and criticism, will retaliate against the person who praises and criticizes.

Begin with interpersonal costliness: arguably, *Costlier* doesn’t originate from interpersonal costs – under the right conditions, third-parties may retaliate in response to either moral praise or moral criticism. Consider examples of third-party retaliation against moral praise:

*Trump voter*

Jerry, who lives in a heavily blue district, thinks Trump supporters are brave, and

says so in front of friends and neighbors. Within a few weeks, word spreads and Jerry loses most of his customers.

*Pro-choice*

Bertha praises Sally's pro-choice views in a heavily pro-life area. As a result, Bertha's neighbors shun her for praising a practice they consider morally abhorrent.

We can easily generate further examples like *Trump voter* and *Pro-choice* where moral praise is interpersonally costly. So, it can be hard to tell whether moral criticism would be *interpersonally* costlier than moral praise – the interpersonal costliness of moral praise and moral criticism varies too much here. For this reason, we should treat the interpersonal costliness of moral praise and moral criticism as on par; this source of costliness doesn't add anything, one way or the other, to evaluating *Costlier*. The subject of moral criticism has better reason to retaliate than the subject of moral praise, *ceteris paribus*. And if interpersonal costliness of moral criticism and moral praise is roughly on par, combined with the asymmetric personal costliness, moral criticism tends to be a costlier signal than moral praise.

In contrast, it is easy to see why moral criticism would often be personally costlier than moral praise: the subject of moral praise lacks reason to retaliate for *being morally praised* – it's nice to be told you're a good person. For example, we shouldn't expect Bob to retaliate against Mary for praising his integrity, even if Bob knows Mary is merely paying him lip service – there is a good chance Mary's moral praise for Bob, even if insincere, may improve his moral reputation. Even if third-parties see such praise as insincere, they will likely disregard it at worst (unless it says something about Bob worthy of moral criticism: e.g., Mary praising Bob for mistreating his pets – moral praise isn't the problem here, but rather what it says about Bob's bad behavior). For this reason, moral criticism tends to be a higher-fidelity signal than praise.

Of course, we should recall the distinction between accountability-criticism and attributability-criticism when thinking about criticism costliness. In the case of *Trump voter*, one criticizing Jerry for his Trump support may not encourage others to punish him, but it may cause Jerry to want to retaliate for damaging his self-conception as a good person, or may dampen his access to the cooperation market. And tweaking the *Trump voter* example somewhat, were someone to criticize Jerry, not only for supporting Trump, but also for casting a vote for Trump in a tight election, then that criticism may encourage punishing Jerry for his actions, and thus encourage Jerry to retaliate in reaction to that punishment. The distinction between different kinds of criticism, and the different sources of costliness they may encourage, applies to Bertha in the *Pro-choice* case too: criticism for her abortion views may harm her self-conception as a good person, and cripple her access to the cooperation market. Had someone criticized her, not just for her pro-choice views, but for successfully lobbying for easier access to abortion services in her state, they may encourage others to punish her, and so risk retaliation that way too. We can see that there are different sources of personal and interpersonal costliness that depend upon whether we're talking about attributability-criticism or accountability-criticism.

Of course, one could improve their moral reputation with insincere criticism. If the payoff were big enough, it may make sense to risk retaliation here. This would threaten to dilute the fidelity of moral criticism as a moral signal: if there are strong enough incentives to insincerely morally criticize, and risk retaliation, *Costlier* may be a bad epistemic heuristic – we couldn't then use incentives to determine the chance that moral criticism and moral praise are sincere. For example, Amy tries to keep her job by

criticizing her co-worker, Jonah, for his insensitive Halloween costume, to improve her moral reputation. Here Amy risks Jonah retaliating, but she may decide the payoff – a boost to her moral reputation – is worth it.

This worry isn't convincing. First, even if there a potentially big payoff here, people are highly risk averse (Weber 1999). While it may be beneficial to insincerely morally criticize others to improve one's moral reputation, people still often refrain. Presumably, this is explained partly by aversion to risk: people greatly value their moral reputation (Vonasch *et al.* 2018), so we should expect a high risk of retaliation in reply to moral criticism, especially when that criticism comes off as hypocritical or self-serving. We should expect that the boost from moral criticism must be sufficiently big, and the risk of retaliation small enough, for people to morally criticize each other, sincerely or not – despite the potential to boost one's moral reputation via moral criticism, it tends to be costlier than moral praise.

Second, the risk is boosted by the empirical evidence that people especially dislike hypocrites: those who criticize others for transgressions of which they are also guilty are seen *as worse* than liars. Add that there is often, though not always, a good chance insincere moral criticism will be exposed as hypocritical or self-serving partly because people are quite good at cheater detection (Lier *et al.* 2013), making this a risky approach. To quote a salient passage:

[People] dislike hypocrites more than direct liars because hypocrites falsely signal. One straightforward explanation for why hypocrites' false signals inspire moral outrage is that misleading other people is generally regarded as wrong ... and hypocrites are especially misleading, because condemnation is an especially convincing signal. A hypocrite's false signals may rouse further disapproval, moreover, because they lead to negative outcomes, such as unfairly boosting the hypocrite's reputation or shaming other people into changing their behavior while the hypocrite carries on. Furthermore, unlike direct statements that one behaves morally, condemnation can harm other people by maligning the condemned – which may make hypocrisy seem particularly wrong. (Jordan *et al.* 2017: 366–7)

Clearly Amy (in the example above) may be somewhat sincere and self-serving in criticizing Jonah's costume. Even here though, she risks being seen by third-parties as hypocritical or self-serving in her moral criticism – in addition to risking Jonah's retaliation, she risks retaliation from third-parties too. There is a real possibility one could use insincere criticism to bolster their moral reputation without suffering retaliation, but the underlying risk acts as a bulwark against doing so which fails to apply to moral praise – insincere moral praise to bolster one's moral reputation only works if third-parties think the praise picks out something praiseworthy, instead of, say, odious. Insincere criticism of others merely to boost one's moral reputation risks reputational damage.

Finally, this worry targets interpersonal costliness: in the example above, Amy wouldn't try to signal her virtuousness, by criticizing Jonah for his insensitive costume, without an audience. So while insincere moral criticism can be risky, Amy may deem it worth the risk with a good chance such criticism improves her moral reputation, and offsets the cost of moral criticism.

The interpersonal costs of moral praise and moral criticism are roughly equal. And if moral praise and moral criticism are roughly equally interpersonally costly, but the moral criticism is personally costlier than moral praise, moral criticism is costlier overall. Since costliness tracks the fidelity of moral praise and moral criticism as signals, *Costlier* is a good epistemic heuristic.



## 4. Two objections

There are a couple of questions remaining for our thesis, and these can be formulated as objections to our thesis. The first objection highlights the gap between establishing that moral criticism is a costly signal, compared with moral praise, but that doesn't show moral criticism is more likely to be *sincere* than moral praise. The second objection is that much of what we've said of moral criticism and praise could (often enough) equally apply to *generic* criticism and praise. But then our thesis isn't specific to *moral* criticism and praise particularly, and this appears to undercut the news value of the thesis; there should be something newsworthy, so to speak, about the claim that criticism is costlier than praise that justifies the interest and application to *moral* criticism and praise specifically.

### 4.1. The sincerity objection

Suppose we've established moral *criticism* is a costlier signal than moral praise. There remains a gap in the argument: just because moral criticism framed as a signal is costlier doesn't by itself show moral criticism is more likely to be *sincere*. We could imagine after all there may be cases where moral criticism is costlier, but people offer moral criticism because the payoff is far larger than the loss threatened by retaliation. So why think *Costlier* is a good enough heuristic to rely on? There are several reasons.

First, moral criticism is often enough (while not always) motivated by emotions. And while it depends on the individual, emotions are often enough difficult to fake (Frank 1988); it is hard to conjure up emotions that aren't organic to motivate criticizing someone. We should expect moral criticism to be tied with emotions as morally criticizing someone can be a risky move – especially as *moral* criticism, unlike much generic criticism, can have a far greater impact on the well-being of the target of that criticism, and so the stakes can be high – we wouldn't normally undertake this unless under the sway of strong emotions. Anger is a good example of this: when in the heat of anger, we may say and do things we would otherwise lack the courage to do.

Second, for signaling systems – e.g. moral criticism – to be robust and stable, they must be hard for signaling free-riders to use that system at a low cost; otherwise, free-riders would use the system, and it would collapse because the low-fidelity signal couldn't be trusted. Since genuine commitment more likely gives rise to a more reliable signal than merely instrumental (Frank 1988), we should expect if moral criticism is a high-fidelity signal, it would be sincere enough; it would be likely here that moral agents who engage in signaling via moral criticism aren't aware of the signaling component of the practice of moral criticism for this reason too.

Third, with respect to practices like moral outrage, and emotionally charged moral criticism, we have experimental evidence that folks are often sincere: researchers asked participants to express their moral outrage at other participants who refused to share money they had been given to be distributed in the course of an experiment (Jordan and Rand 2020). They found that among outraged participants, outrage decreased once they were given the opportunity to contribute to others. That participants who had the chance to share felt less outrage is evidence that expressing moral criticism of others, like moral outrage, has as one of its major functions to signal sincere commitment: the opportunity to share with others affords a better opportunity to signal their sincere moral views, and this greatly diminished their need to use a signal like moral outrage as a means of reputational management. This evidence, when taken together, is at least very good evidence that moral criticism is a high-fidelity signal compared with moral praise.

#### 4.2. *The generic objection*

Perhaps you're convinced moral criticism is costlier, and more likely sincere, than moral praise. However, you may think this just as easily applies to *generic* criticism and praise, and that isn't interesting where moral philosophy is concerned. However, we think this reaction is misguided. There are a few reasons moral philosophy should be especially interested in *Costlier*:

First, sincere moral feedback, like moral criticism, can provide useful feedback to facilitate our improvement as moral agents. We can clearly mature as moral agents from feedback we receive from those in our social circles who know us quite well, but we shouldn't neglect feedback from strangers who, though they may not know us well, are in a better epistemic position to see our faults as moral agents than those who care about us; there no doubt can be a kind of objectivity that results from a certain sort of epistemic and personal distance. We assume that progressing as moral agents has moral (and other kinds of) value; so moral feedback, like moral criticism, would likewise have moral value.

Second, sincere moral feedback, like moral criticism, is useful information both as to how others see us, and the kind of moral agents we are. We may be unaware that others view us in a certain way – as is often enough the case – and we need something like moral criticism to help create a better representation of how we're seen by others. Sincere moral criticism from strangers can be valuable feedback about us as moral agents partly because third-parties won't have as much at stake in our moral evaluations. We may distrust the moral evaluations of loved ones because they may hold back out of concern for our feelings, or from people in our circles who don't like us because they may be out to hurt us without good reason. Strangers potentially offer feedback that isn't as likely to be tainted by prior involvement, and so provide valuable knowledge about our moral character and tendencies (especially if the moral criticism from strangers is robust; i.e., we find we receive the same criticism, independently, from different strangers).

Finally, sincere moral feedback can modulate our behavior for the better or worse; if we think third-parties *really* think we're acting like a jerk, we have reason to reform our ways not only because we (likely) don't want to be jerks, or be seen as jerks, but because this criticism may be a warning that retaliation is in the offing if we don't reform our ways. Not only would that be bad for our projects and interests, but it may be the impetus to reform our behavior to improve our relationships with people we care about; we can use sincere moral criticism from strangers to improve relationships with friends and loved ones. And surely these are reasons to think that *Costlier* is particularly valuable in the moral domain.

### 5. Conclusion

Let us review. Treating moral praise and criticism as signals with varying degrees of costliness can illuminate their respective fidelity. And so, we saw moral criticism is often a costlier moral signal, with more potential to incite retaliation, than moral praise. Thus we should expect one to risk retaliation from moral criticism, especially in cases where they believe someone *deserves* to be criticized, and perhaps not even then. And though much of the research we've covered in support of *Costlier* shows that criticism, moral and non-moral alike, is likely costlier than praise, the thesis should be of particular interest in the moral domain for several reasons: it (i) offers an insight into how others actually view us as moral agents; (ii) is feedback to help us improve our moral characters; and (iii) can encourage some behaviors, and discourage others. These things

can enhance our relationships with others and refine our moral agency. Since moral criticism and moral praise are inputs that influence moral reputations, facilitate valuable cooperation, and can help us better shape our moral agency, *Costlier* is an important moral heuristic.<sup>1</sup>

## References

- Bird R. and Smith E.** (2005). 'Signaling Theory, Strategic Interaction, and Symbolic Capital.' *Current Anthropology* **46**, 221–48.
- Boyd R., Gintis H., Bowles S. and Richerson P.J.** (2003). 'The Evolution of Altruistic Punishment.' *Proceedings of the National Academy of Sciences USA* **100**, 3531–5.
- Cowen T.** (2012). *An Economist Gets Lunch*. London: Dutton Books.
- Dibble J.L. and Levine T.R.** (2013). 'Sharing Good and Bad News with Friends and Strangers: Reasons for and Communication Behaviors Associated with the MUM Effect.' *Communication Studies* **64**, 431–52.
- Frank R.H.** (1988). *Passions Within Reason: The Strategic Role of Emotions*. New York, NY: Norton.
- Fu G., Heyman G.D., Qian M., Guo T. and Lee K.** (2016). 'Young Children With a Positive Reputation to Maintain Are Less Likely to Cheat.' *Developmental Science* **19**, 275–83.
- Hardy S.A. and Carlo G.** (2011). 'Moral Identity: What Is It, How Does It Develop, and Is It Linked to Moral Action?' *Child Development Perspectives* **5**, 212–18.
- Henrich J.** (2015). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.
- Henrich J., Heine S.J. and Norenzayan A.** (2010). 'The Weirdest People in the World?' *Behavioral and Brain Sciences* **33**(2–3), 61–83.
- Jordan J.J. and Rand D.G.** (2020). 'Signaling When No One is Watching: A Reputation Heuristics Account of Outrage and Punishment in One-Shot Anonymous Interactions.' *Journal of Personality and Social Psychology* **118**, 57–88.
- Jordan J.J., Hoffman M., Bloom P. and Rand D.G.** (2016). 'Third-Party Punishment as a Costly Signal of Trustworthiness.' *Nature* **530**, 473–6.
- Jordan J.J., Sommers R., Bloom P. and Rand D.G.** (2017). 'Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling.' *Psychological Science* **28**, 356–68.
- Kogelmann B. and Wallace R.H.** (2018). 'Moral Diversity and Moral Responsibility.' *Journal of the American Philosophical Association* **4**, 371–89.
- Lier J.V., Revilin R. and Neys W.D.** (2013). 'Detecting Cheaters without Thinking: Testing the Automaticity of the Cheater Detection Module.' *PLoS ONE* **8**(1), e53827.
- Miller G.F.** (2007). 'Sexual Selection for Moral Virtues.' *Quarterly Review of Biology* **82**, 97–125.
- Moller D.** (2013). 'The Epistemology of Popularity and Incentives.' *Thought: A Journal of Philosophy* **2**, 148–56.
- Rosen S. and Tesser A.** (1970). 'On Reluctance to Communicate Undesirable Information: The MUM Effect.' *Sociometry* **33**, 253–63.
- Shoemaker D. and Vargas M.** (2019). 'Moral Torch Fishing: A Signaling Theory of Blame.' *Noûs*. <https://doi.org/10.1111/nous.12316>.
- Simler K. and Hanson R.** (2018). *The Elephant in the Brain: Hidden Motives in Everyday Life*. Oxford: Oxford University Press.
- Spence M.** (1973). 'Job Market Signaling.' *Quarterly Journal of Economics* **87**, 355–74.
- Sperber D. and Baumard N.** (2012). 'Moral Reputation: An Evolutionary and Cognitive Perspective.' *Mind and Language* **27**, 495–518.
- Vonasch A.J., Reynolds T., Winegard B.M. and Baumeister R.F.** (2018). 'Death Before Dishonor: Incurring Costs to Protect Moral Reputation.' *Social Psychology and Personality Science* **9**, 604–18.
- Watson G.** (1996). 'Two Faces of Responsibility.' *Philosophical Topics* **24**, 227–48.
- Weber E.U.** (1999). 'Who's Afraid of a Little Risk? New Evidence for General Risk Aversion.' In J. Shanteau, B.A. Mellers and D.A. Schum (eds), *Decision Science and Technology*, pp. 53–64. New York, NY: Springer.

<sup>1</sup>Thanks to an anonymous referee whose comments made the paper substantially better. And even without anonymity, in this case, criticism *doesn't* risk retaliation.

- Weenig M.W.H., Groenenboom A.C.W.J. and Wilke H.A.M.** (2001). 'Bad News Transmission as a Function of the Definitiveness of Consequences and the Relationship Between Communicator and Recipient.' *Journal of Personality and Social Psychology* **80**(3), 449–61.
- Yamagishi T.** (1986). 'The Provision of a Sanctioning System as a Public Good.' *Journal of Personality and Social Psychology* **51**, 110–16.
- Zahavi A** (1975). 'Mate Selection – A Selection For a Handicap.' *Journal of Theoretical Biology* **53**, 205–14.

**Jimmy Alfonso Licon** is an Emergent Ventures Fellow at the Mercatus Center at George Mason University. He works on issues in ethics, epistemology, and philosophy, politics, and economics (PPE). He holds a doctorate in philosophy from the University of Maryland.