Evolution of moral outrage: I'll punish your bad behavior to make me look good

February 25 2016, by Jillian Jordan, Yale University

Standing up for what's right can come with a cost to the individual – but also a benefit. Credit: Michael Fleshman, CC BY-NC

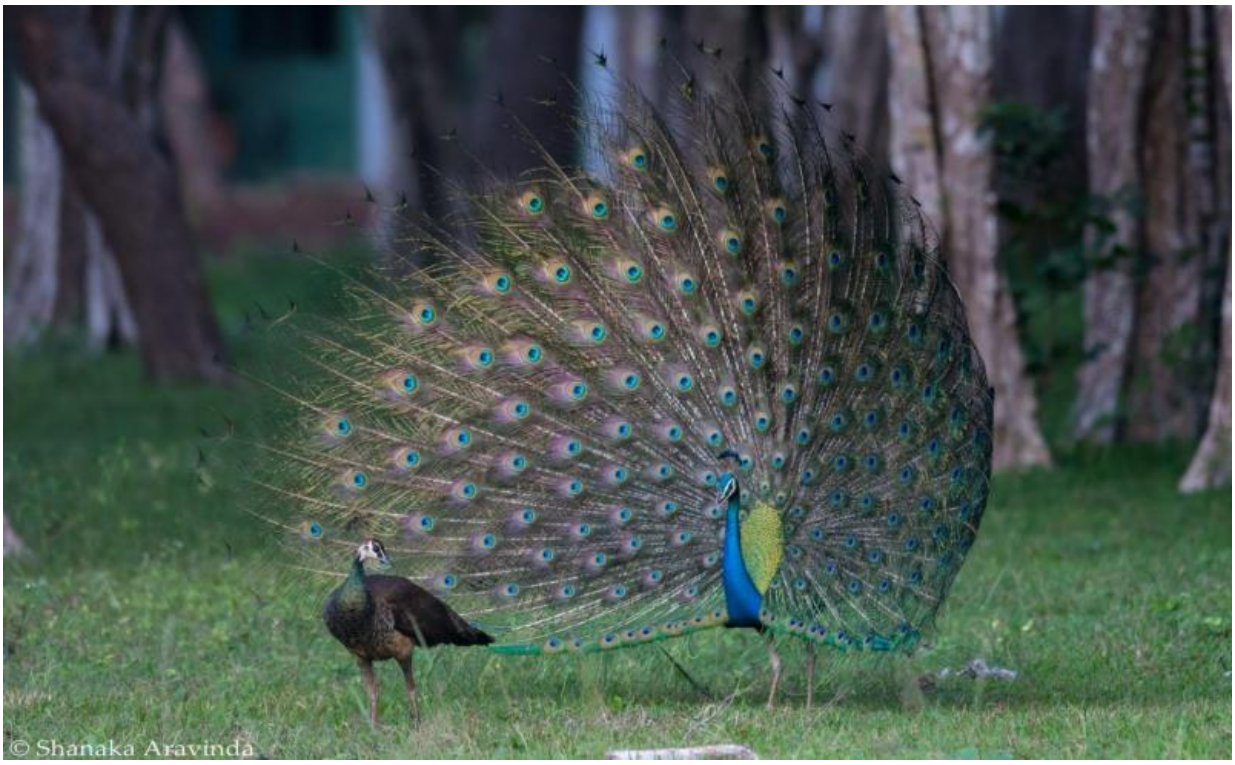
What makes human morality unique? One important answer is that we care when other people are harmed. While many animals retaliate when directly mistreated, humans also get outraged at transgressions against others. And this outrage drives us to protest injustice, boycott companies, blow whistles and cut ties with unethical friends and colleagues.

Scientists refer to these behaviors as third-party punishment, and they have long been a mystery from the perspective of evolution and rational self-interest. Why should people invest time, effort and resources in punishing – even when they haven't been harmed directly? While it's clear that our punishment is motivated by moral outrage, that raises the question of why we developed a psychology of outrage in the first place.

Why punish, since it comes with a cost?

One theory is that people punish to benefit society. Social sanctions from peers can deter misbehavior, just as legal punishment does. To take an example from daily life, if Ted decides to criticize his coworker Dan for going on Facebook during work, Dan and others will be less likely to slack off, and the company will be more productive. Perhaps, then, Ted punishes Dan to promote a successful workplace.

However, this logic can fall prey to the "free-rider problem": everyone wants to be at a successful company, but nobody wants to sacrifice for it. If Ted punishes Dan, Dan might exclude him from his upcoming party. Why should Ted take this hit?



Have a look at my plumage; you know what this dazzling display means. Credit: Shanaka Aravinda, CC BY-NC-ND

One reason individuals might benefit from punishing is via rewards for deterring misbehavior: Dan's boss might reward him for promoting company productivity by criticizing Ted.

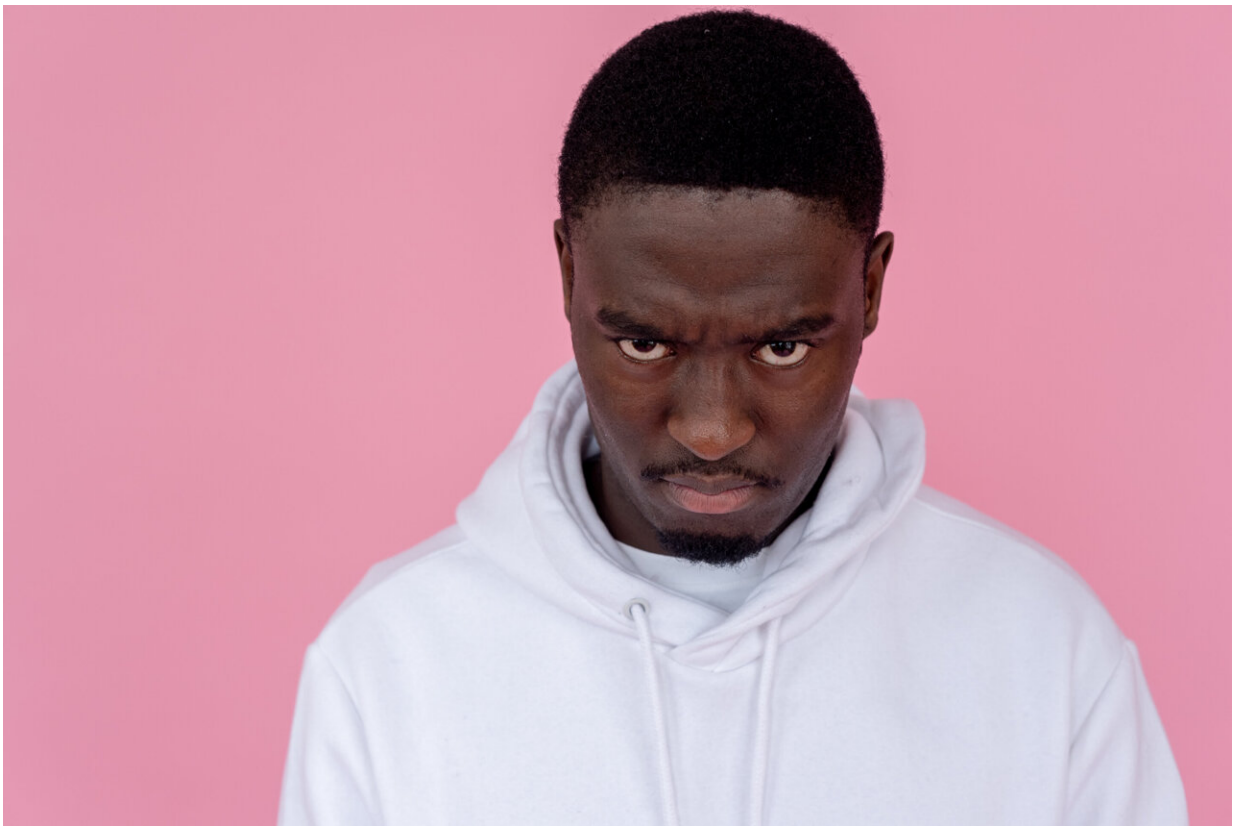
In recent Nature paper, my colleagues and I provide evidence for a different theory of individual benefits of punishment – one that can operate in conjunction with the rewarding process described above. We argue that individuals who punish can boost their reputations by signaling that they can be trusted. If Dan punishes Ted for going on Facebook, his other coworker, Charlotte, might trust that he won't slack off if assigned to an important project.

Signaling one thing by doing another

To make our case, we first created a game theory model of third-party punishment as a "costly signal" of trustworthiness.

The concept of costly signaling originated with the example of the peacock's tail. Female peacocks want to mate with males who have good genes, but they cannot directly observe genetic quality. So high-quality males woo females with elaborate plumage, which they can afford to produce only because they have good genes. It's too energetically expensive for low-quality males to produce the same kind of beautiful tails; the cost of trying to do so would be enormous, and not worth the benefit of attracting mates by (falsely) appearing to be high-quality. So beautiful tails end up being a reliable signal for genetic quality. (The same logic can be applied to people signaling their wealth with extravagant watches or sports cars.)

Our model is based on the idea that, just as peacocks vary in their genetic quality, people vary in their incentives to be trustworthy. Imagine that Ted and Eric are both summer interns. Ted aspires to work at the company in the long run, while Eric just wants to add a line to his resume. Both Ted and Eric wish to be selected by Charlotte for the aforementioned project (as getting selected means getting paid more), but they will behave differently if selected. Ted has the incentive to work hard – even at the cost of his weekend plans – because doing so will boost his career prospects in the company. In contrast, Eric will get the line on his resume regardless of if he does a good job, so his incentive is to slack off and enjoy his weekend.



Credit: SHVETS production from Pexels

In situations like this, people like Charlotte (whom we call Choosers in our model) must decide whether to trust people like Ted and Eric (whom we call Signalers) – who are either trustworthy (like Ted) or exploitative (like Eric). Choosers cannot directly tell who is trustworthy – if Charlotte asked Eric whether he would work hard, he would say yes: he wants to get the raise! Thus, Choosers must base their decisions on costly signals. Can third-party punishment be such a signal?

We argue that the answer is yes, because the same factors that motivate people to be trustworthy often also motivate them to deter misbehavior via punishment. For example, Ted's drive to get ahead in the company

gives him an incentive to be trustworthy to Charlotte – and also to get rewarded by his boss for punishing Dan. Consequently, the benefit of impressing Charlotte, when combined with the reward from his boss, could be enough to outweigh the cost of punishing.

In contrast, because Eric doesn't value a reward from his boss very much, he might not find it worth punishing Dan to impress Charlotte. As a result, punishment can serve as an honest and reliable signal of trustworthiness.

From theory to data: economic experiments on how people punish

Next, we tested this theory using incentivized experiments where we had human subjects engage in a stylized version of the scenario described above. In our experiments, a Signaler subject had the opportunity to sacrifice money to punish a stranger who had treated somebody else selfishly. Then in a second stage, a Chooser subject decided whether to entrust the Signaler with some money – and then the Signaler got to decide how much of the money to return.

The results? As predicted, Choosers were more likely to trust Signalers who had punished selfishness in the first stage. And they were right to do so: Signalers who punished really *were* more trustworthy, returning more money in the game. Furthermore, when Signalers had a more direct way to signal their trustworthiness to Choosers (by sharing money with a stranger, rather than punishing somebody for not sharing), they were less likely to punish – and Choosers were less likely to care whether they did.

Implications for human morality

Thus, we provide evidence that punishing selfishness can act like a

peacock's tail – it can serve as a public display that hints at a quality (trustworthiness) that can't easily be observed. We help resolve the "free-rider" problem by showing that individuals who punish others benefit from an improved reputation. And we help explain why we might have developed a sense of moral outrage in the first place.

Our theory can also speak to why people sometimes punish wrongdoing that could *never* affect them personally, even in the future. For example, why do men condemn sexism, even though they have no personal stake in wiping it out? One explanation may be to signal to women that they can be trusted not to behave in a sexist manner.

The signaling account can also help explain our fiery hatred of hypocrites who punish others for behaviors they engage in themselves. Such hatred seems strange when you consider that punishment can help society by deterring misbehavior – if you're going to behave badly yourself, isn't it better to at least chip in by punishing wrongdoing? Yet we think hypocrites are much more contemptible than people who behave badly but do not punish others. This perspective makes sense when you consider that hypocrites engage in dishonest signaling – their punishment falsely advertises to others that they can be trusted.

Finally, our theory sheds light on when punishment does – and doesn't – benefit the group and society. Punishment generally deters misbehavior: when Ted punishes Dan to impress Charlotte and get rewarded by his boss, he is likely to improve workplace productivity. But people don't always punish in the ways that are best for society. Ted may face similar incentives to punish Dan even if Dan has already been punished by others – or if Ted (but only Ted) knows that Dan's perceived transgression was actually a well-intentioned mistake. Thus, people may engage in disproportionate punishment, or punish accidents, for the purpose of boosting their own reputations. These examples demonstrate that if punishment evolves to benefit individuals, we should expect

imperfect outcomes for society when individual and collective incentives do not align.

Moral outrage and third-party punishment are key features of human morality, and set us apart from other animals. Our research suggests that the drive to punish has a self-interested side, and may exist, in part, to boost our reputations. This conclusion doesn't undermine the moral good that often results from our drive to punish, but rather sheds light on its origins and its nature.

More information: Jillian J. Jordan et al. Third-party punishment as a costly signal of trustworthiness, *Nature* (2016). DOI: 10.1038/nature16981

This article was originally published on The Conversation. *Read the* original article.

Source: The Conversation