The side-taking hypothesis for moral judgment
Peter DeScioli

A recent theory proposes that moral judgment is an evolved strategy for choosing sides in conflicts. Evidence from moral psychology undermines previous evolutionary theories of morality focused on cooperation. I show how the side-taking hypothesis explains these otherwise puzzling patterns of moral judgment — especially its focus on actions rather than intended consequences. Moral judgment offers an alternative to choosing sides based on status or relationships by conditioning support on disputants' actions rather than their identities. Observers can use moral judgment to take the same side, avoiding costly fights among themselves, while dynamically changing who they support in different conflicts. Observers can benefit from moral side-taking even when condemnation harms other individuals and society.

Address
Stony Brook University, Stony Brook, NY, United States

Corresponding author: DeScioli, Peter (pdescioli@gmail.com)

Imagine you are alone on a desert island when a hulking man says to hand over your food supplies or he will kill you. You protest that his threats are morally wrong. How much protection could your moral judgment really provide? If you had to choose, would you defend yourself with a hand axe or moral arguments? Now imagine instead that there are fifty people on the island. In this case, morality might actually be the better weapon: a persuasive moral argument could rally dozens of armed defenders to your side.

A recent theory begins to uncover how something so seemingly feeble as moral judgment could possibly offer real advantages in evolutionary competition. DeScioli and Kurzban [1••] argued that this puzzle can be resolved by thinking about how humans choose sides in conflicts. A persuasive moral argument can launch a barrage of stones at an opponent. It can sever an opponent's relationships, cut off trade and supplies, and cast them into exile. However, these powers require an audience who will listen to moral accusations. Hence, a critical challenge is figuring out why observers care about who is right and wrong, and why they can sometimes be persuaded to oppose wrongdoers, even their own family and friends.

Explaining moral judgment, not cooperation
Traditional evolutionary theories of morality focus on understanding cooperation [2–4]. These theories do an excellent job of explaining why humans and other animals care for offspring, cooperate in groups, trade favors, communicate honestly, and respect property. However, they do not explain moral judgment — why humans compute the wrongness of actions, announce moral judgments, condemn other people's violations, and debate over moral rules. For example, kin selection explains why many animals care for their own offspring; it does not explain why humans morally judge other parents who neglect offspring — an unusual behavior absent in other species.

Research from moral psychology reveals the information-processing structure of moral judgment (reviewed in [5]), which is the primary evidence for distinguishing its evolved functions [6]. Moral judgment takes as input a person's action and computes as output a wrongness value for that action. People then announce their moral evaluations, argue about the wrongness of actions, and show hostility toward wrongdoers.

Moral judgment's focus on actions is among its most distinctive and mysterious features. Similar to Kant's deontological philosophy, people often judge actions such as lying, theft, killing, premarital sex, or blasphemy to be morally wrong even when these actions can achieve better outcomes [5,7,8]. A well-known example is people's judgments that it is morally wrong to kill one person to save five people in the footbridge trolley problem [9]. Deontological judgment is puzzling because evolved decision mechanisms are usually consequentialist in that they evaluate the outcomes of actions rather than the actions themselves. For example, people's precautionary judgments aim to avoid injury (an outcome) rather than judging skydiving or bullfighting (actions) to be inherently injurious.

Why then do humans assign moral values to actions? Moral judgment is not required for cooperation: Many non-human organisms cooperate while few, perhaps none, assign moral values to actions. Neither is moral judgment a particularly effective executor of cooperation: Research shows that people frequently behave selfishly even when they judge the same behavior to be immoral

[10]. Another idea is that moral judgment enforces co-operation by punishing cheaters. But this theory too predicts consequentialist mechanisms, specifically punishment aimed at deterring harm. Instead, research shows that people's moralistic punishments are aimed at retribution rather than deterrence [11], and its targets include harmless and beneficial behaviors such as same-sex marriage, birth control, interest-bearing loans, and scientific inquiry.

In fact, moral judgment can be extremely destructive: moral condemnation reduces economic efficiency [12], incites political extremism [13**,14], damages close relationships [15], radicalizes terrorists [16], motivates sexual-orientation hate crimes [17], and blocks medical services for pregnant women, people addicted to drugs, and HIV patients [18,19].

A common fallback for cooperation-based theories is to attribute evidence inconsistent with cooperation to mistakes in moral cognition due to erroneous information [20] or simple heuristics [21,22]. These proposals stand on weak theoretical ground because people's deontological ''mistakes'' are large in magnitude, highly structured, and absent in non-human animals as well as other types of human judgments (e.g., welfare, precautions, economics). They also stand on weak empirical ground. Haidt [5] tested and falsified a mistake hypothesis by finding that people condemn actions that even they themselves report to cause little or no harm. Kurzban *et al.* [23*] tested and falsified a mistake hypothesis by finding that participants reported greater willingness to kill a brother to save five brothers, compared to killing a stranger to save five strangers, contradicting the claim that deontic resistance to killing is driven by altruism heuristics. DeScioli *et al.* [24] tested and falsified a mistake hypothesis by finding that people are more likely to choose harmful omissions when there is a threat of punishment, contradicting the claim that people's greater willingness to harm by omission than commission results from cognitive errors (see also [25,26]).

In short, theories of cooperation explain many features of human social life. If humans were Vulcans from *Star Trek*, these theories might also explain morality. However, the data show that humans are hybrid Vulcan-Kantians whose moral judgments focus not only on intended outcomes but also on the particular actions people choose. Our Kantian side formulates moral rules of action, argues about the rules, and creates the diversity of moral systems across cultures. It is the aspect of morality that is most distinctively human and challenging to understand.

Is morality mightier than the sword?

There is little evidence in other animal species of moral judgment, moral announcements, or moral accusations, much less moral debates. Non-human animals do not argue about the morality of sharing food, paying debts, same-sex relationships, or military invasions, even when they exhibit these behaviors. To understand such an unusual trait, it might help to consider other distinctive features of human social ecology. One possibility is human conflict.

Conflict is much more complicated in humans than most animals because people not only support each other in fights, but also switch sides [27]. In most animals, fighting is largely a matter of assessing the power of a single opponent and knowing when to back down [28,29*], but assessment is much more difficult when the opponent can recruit help from other individuals [30]. An opponent who initially appears weak can suddenly multiply in strength with the arrival of a crowd of supporters. This complexity applies not only to fistfights and arrow volleys but also to verbal arguments and gossip campaigns.

In the arena of human conflict, moral judgment is a powerful weapon. Accusations of wrongdoing such as lying, infidelity, or blasphemy can mobilize a mob of aggressors against an opponent. Moral condemnation turns cold shoulders toward colleagues, sends terrorists after political cartoonists, and pulls nations into costly wars.

In this social environment, it is easy to see why natural selection would favor psychological adaptations for using moral judgment against opponents, and for avoiding prohibited actions to escape the wrath of condemners. The difficulty is understanding why an audience cares about right and wrong, and why their punishment is directed at a diverse and changing list of particular actions.

Choosing sides

The audience members observing a human conflict face their own challenges. Humans live in dense social networks and individuals often have cross-cutting ties to both sides of a conflict. When both sides request support from the same person, that person will inevitably damage at least one relationship, creating a difficult choice. Further, to assess the consequences of their side-taking decisions, observers also need to evaluate how other observers will choose sides.

One side-taking strategy is bandwagoning — supporting the higher-status against lower-status fighter. This tactic curries favor with the more powerful disputant and joins with others using the same strategy. However, bandwagoning can set a dangerous precedent: higher-status individuals learn that they can easily exploit lower-status individuals, creating a threat to everyone except the top of the hierarchy.

A second side-taking strategy is alliance-building — supporting one's own supporters [27,31]. By teaming