Annals of the New York Academy of Sciences

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

ESSAY

The Moral Sword

Peter DeScioli

Department of Political Science, Stony Brook University, Stony Brook, New York, USA

Correspondence: Peter DeScioli (pdescioli@gmail.com)

Received: 29 April 2025 | **Revised:** 18 July 2025 | **Accepted:** 27 August 2025

Funding: The author received no specific funding for this work.

Keywords: alliances | benevolence | condemnation | evolutionary psychology | moral judgment

ABSTRACT

Why do we hurt people in the name of morality? Here we elaborate on the theory that moral judgment is an evolutionary strategy for choosing sides in conflicts. Hurting wrongdoers is part of the strategy. Morality may seem like a guiding light for cooperation, but it is actually closely tied to aggression. As a result, moral condemnation is not always good: It is a gamble that risks punishing the innocent and inflaming hostilities between factions. These dangers are obscured by confusing morality with benevolence. Thus, we examine how moral judgment fundamentally differs from benevolence and goodness. The argument appeals to evolutionary psychology, moral psychology, and what we will call the method of natural language. Accordingly, we will minimize jargon and scholarly accounting to address a general audience across the many disciplines concerned with morality. The final section provides a concise review of the essential literatures for further reading.

1 | Introduction

We may think of morality as a force for good. But what morality protects—life, liberty, property—it can also destroy. What explains the dark side, the fury, the flames, the lashing, the killing, the inquisitions, the persecutions, the witch hunts, all conducted under moral pretenses? If morality is good, then why do we hurt people in its name? The reason, I think, is that morality is not the perfect angel it seems to be. Morality comes from moral judgment, and moral judgment is a human strategy for conflict. Violence, for good and for bad, is part of its nature. Let me explain.

Morality is commonly held in the highest regard as the essential foundation of society and further still as an infallible compass, as the source of all value and meaning, and as the sure path to heaven. For example, we see this sentiment implied when people constantly complain that morality is in decline. In yearly surveys over the past two decades, 40%–50% of Americans rated the country's moral values as currently "poor" [1]. Indeed, psychologists have argued that the perception of moral decline is an enduring illusion found widely across times and cultures [2]. Many people feel that morality is always in short supply.

On the other hand, skeptics see in morality many of the worst impulses in human nature. Moral condemnation unleashes sadistic motives to make wrongdoers suffer pain, humiliation, butchery, and execution. A long history of public flogging, branding, mutilation, and decapitation attests to its brutality. And we see many more horrors, from mind control and censorship to riots and terrorist attacks. The skeptical view is captured by a memorable line from the philosopher Bertrand Russell: "The infliction of cruelty with a good conscience is a delight to moralists. That is why they invented Hell" (p. 5) [3]. According to skeptics, we have too much morality, not too little.

We are here to understand the dark side: why moral judgment motivates condemnation, aggression, and at times, cruelty and destruction. We will begin with a simple answer carried by the metaphor of moral roulette. Moral condemnation is a gamble

that risks punishing the innocent, so the dark side is inherent in the moral game. Then we turn to the confusions that animate the question, considering why the dark side may not be obvious, why morality wears a halo. The halo may come from confusing morality with benevolence, which is actually a different mental faculty and a more reliable source of good. With the halo removed, our view of moral judgment returns to Earth, and its ties to aggression come into focus. Even so, we argue that moral aggression serves a critical function in resolving conflicts, limiting the damage from three worse sources of aggression: selfishness, hierarchy, and alliances. Furthermore, this account suggests that moral aggression turns destructive when people use it to instigate conflicts instead of to settle them.

Having seen how confusion obscures and encourages moral aggression, we then set out to clarify the distinctions between goodness, benevolence, and morality. We summarize three essential tools: evolutionary psychology, moral psychology, and the method of natural language. We then apply these tools to reveal how morality differs from benevolence and goodness.

2 | Moral Roulette

Moral judgment is fundamentally a director of aggression. Lady Justice weighs the scale to ultimately direct her sword. When we judge that someone's actions are morally wrong, we feel angry toward them. Anger prepares us to fight. It inhibits our sympathy and makes us want to hurt the offender. We contemplate ways to injure their body, their reputation, and their relationships. We spread news of their crimes to incite an enraged crowd against them. Moral reasoning can be complex, refined, creative, and earnest, but it eventually ends with the sword: threats, insults, smears, screams, punches, riots, pitchforks, and guillotines.

What if the accused is not guilty? What if their offense is not a genuine crime? What if the punishment does more harm than good? Then moralistic aggression will destroy, not protect, innocent lives.

Even a slight miscalculation can turn justice into injustice. Execute the wrong person for murder and you become the murderer. Seize goods from a mistaken thief and you become the robber. Mistake the truth for a lie and you become the liar.

If you hope to do good, the pursuit of moral justice is a risky gamble, playing roulette with other people's lives. If you punish the guilty fairly, then you win and they lose. If you punish the innocent or punish too severely, then everyone loses.

Moral judgment, then, is inherently aggressive, and its inevitable errors, moral misjudgments, inherently cause injury and injustice. Moral judgment is a gamble and gamblers often lose.

Thus, the dark side of morality is not an anomaly or exception, not a distortion or well-meaning excess. Rather, it is the rule of our moral nature, the necessary cost of playing moral games. Spin the moral wheel, round and round it goes, to virtue or to vice, nobody knows.

3 | Not the Best and Not the Worst of Human Nature

Now, if you find yourself surprised by moralistic aggression, then you have probably confused moral judgment with its gentle cousin, benevolence [4]. Benevolence, compassion, altruism, cooperation, generosity, charity, love, kindness: Our friendly motives are a safe bet for doing good. Lady Benevolence does not have a sword. She carries a basket to gather fruit, which she gives out to everyone. She creates abundance and prosperity. She does good consistently, since you cannot go too wrong giving out fruit. Benevolence knows no anger, casts no stones. She is not the same as moral judgment. She has a mind of her own.

Morality is not gentle and kind like benevolence. It is not the best of human nature. It is not a reliable guide to life in general.

But neither is moral judgment the worst source of aggression, and that is why we admire it. The worst aggression comes from three different strategies: selfishness, hierarchy, and alliances. These menaces have no blindfold or scale, no regard for law, impartiality, evidence, or truth. Selfishness lies, steals, and kills to get its way. Hierarchy dominates and exploits. Alliances incite endless cycles of vengeance between us and them. Alliances inflict damage exponentially because they recruit more and more fighters to the fray, multiplying the injuries to everyone in round after round of retaliation.

Humans, therefore, live in a treacherous world of conflict, dominance, and alliances. In this dangerous world, moral judgment promises refuge in impartial laws and measured aggression [5–7]. Lady Justice stands between warring factions, holding enemies at bay with reason, truth, and principle. She aims above all to end the conflict, to settle the dispute, to restore peace. She brings down her sword on the guilty so that enemies may put their swords down.

When a conflict erupts, we can use moral judgment to direct aggression toward ending the fight instead of escalating it, toward concluding the violence instead of inciting more of it. To choose the target of aggression, moral judgment weighs the actions that the opponents have taken. Whoever has taken the most wrongful action is judged the wrongdoer who deserves punishment. If most people judge the opponents' actions by the same moral code, then they can coordinate their aggression against the same opponent, instead of dividing by loyalty into opposing factions. By dealing punishment fairly, impartially, and proportionally, moral judgment aims to appease most observers and allies on all sides. When it succeeds—and often it does not—moral aggression delivers a decisive blow, settles the conflict, and stops the volley of attacks between rival alliances.

Moral judgment settles conflicts by punishing offenders for their *actions*. In contrast, hierarchy and alliances motivate us to punish *persons*, the person who is subordinate in the hierarchy, or the person from a rival alliance. Punishing persons provokes more aggression. Punishing the subordinate fighter encourages the dominant fighter to exploit subordinates again. Punishing the fighter from a rival alliance provokes their allies to retaliate in kind. Thus, punishing persons is generally more incendiary

and destructive than punishing actions—the strategy of moral judgment.

Moral judgment risks punishing the innocent, but taking that gamble is better than the certainty of continuing the cycle of retaliation. Even if you mistakenly condemn the innocent, as long as most observers agree, punishing them can still signal the end of the conflict and spare everyone from further attacks. Punishing the innocent does not do good—everyone still loses—but it can nonetheless limit the losses relative to continued retaliation, enduring a lesser tragedy to prevent the worst. To settle the fight, the evidence of guilt must be credible enough to persuade most people, including neutral parties and, ideally, the defendant's allies—in other words, beyond a reasonable doubt. If most people agree, then moral judgment can settle a conflict successfully, even when they have mistakenly condemned the innocent. These are the brutal calculations of moral roulette, prudent and practical to be sure, but hardly the work of an angel.

We can now say precisely what is so wonderful, so admirable about morality: Moral judgment helps us settle conflicts. Humans live in a minefield of rival factions, and moral judgment lights the way to peace. We invoke moral laws like magic words, *thou shalt not kill*, to unite rival factions against a wrongdoer. Under the siege of constant war, what could be more beautiful than the moral laws that promise peace?

4 | When Moralists Strike First

Conflict, then, is where moral judgment serves as an indispensable strategy for peace and security. In conflict, morality is the lesser evil, less evil than selfishness, hierarchy, and alliances. When we are locked in conflict, we need moral judgment to escape. That is when morality shines and bestows its virtues, not through benevolence and kindness, but by settling a fight with a final blow of fair punishment.

In peace, however, moral judgment is useless and troublesome, an agitator and an instigator of conflict. What use is the sword of justice when no one has raised a hand against another? Suppose everyone is walking merrily through a lush valley, gathering fruit in baskets and sharing it freely. How out of place would Lady Justice be with her scale and sword? No one needs her laws, judgment, or punishment. In this happy place, poking around with a sword can only do mischief.

Consider what may happen when Lady Justice becomes an instigating moralist. She interrupts the amiable company with *thou shalt not* do this, or say that, or go here, or use this thing that way, for these actions are wrong, and *thou must* do this and that, here and there, for these actions duty compels. She starts to condemn harmless actions such as songs, jokes, and taboo ideas like human evolution and the origin of the universe. Unbidden and unwanted, she hunts for wrongdoers though no one has been wronged.

Soon enough, someone else picks up a sword in defense after being needlessly poked and prodded with *shalt not, must*, and various moral accusations. Then the others see the trouble brewing, take up arms, and call on their allies. The community divides into factions, and each alliance plots to strike first before their enemies do, trapped in a dilemma of surprise attack. They are now in full-blown conflict, with hostilities raised beyond moral condemnation to divisive factions. It all began with someone using moral judgment out of place, to start a conflict instead of to end one.

Moral judgment can help settle conflicts, but it is not good for much else. Outside of conflict, moral condemnation amounts to an unprovoked threat of aggression. Moral imperatives bearing *shalt not* and *must* inherently carry a threat of coordinated attack. When a moralist, for instance, warns a speaker, *Thou shalt not take the name of the Lord thy God in vain* (echoing the Biblical commandment), they are not expressing kindness or heavenly virtue; rather, they are threatening the speaker with the wrath of earthly punishment, including the terror of eternal hellfire. To condemn without provocation is to start a fight. Starting fights is bad whether your motives are moral or selfish. Indeed, moral condemnation is often just a pretense for selfish aggression, a cover the aggressor uses to recruit allies in their attack.

In addition to starting trouble, moral judgment is useless for most of life's challenges, apart from settling conflict. Moral judgment cannot help gather food, catch fish, or farm the land. It cannot build shelters, invent tools, or understand nature. It cannot heat homes or protect us from storms and floods. It cannot lay roads, build cities, or supply water and electricity. Moral judgment does not provide the greatest happiness to the greatest number, as Bentham wished it would. Even in social life, morality's uses are limited. You need more than moral imperatives to attract a mate, care for a child, and sustain lasting friendships. Productive cooperation requires skills, planning, and teamwork, not just moral laws like *thou shalt not cheat*.

These limits are obvious but overlooked by those who see morality as the primary source of good and prosperity, and by moral crusaders who react to every problem by looking for sinners to punish. When food, housing, and medicine are scarce, moralists look for wrongdoers to prosecute instead of producers to supply what is needed. But punishing farmers cannot produce more food, and when the moralists are mistaken, as they commonly are, they only compound scarcity with unjust aggression that deters suppliers. Whether landlords, bankers, corporations, police, journalists, or despised "capitalists," the essential workers who provide for society become essential scapegoats. They are to be ritually sacrificed by a priesthood of moralists who preach that the sword of justice provides everything we need.

As before, once a moralist casts the first stone, alliances mobilize on all sides while suspicions and hostilities rise. The moralist has conjured a real conflict out of nothing. Perversely, now the society may actually need to punish a sacrificial sinner in a show of justice to signal the end of fighting.

When moralists strike first, therefore, they start fights and add divisive aggression to society's problems. Unprovoked condemnation is another form of moral misjudgment, one of the errors that lead moralists to hurt people needlessly. In addition to facts, laws, and punishment, we can also be mistaken about when moral judgment is beneficial or harmful: condemning

someone's actions when there is no immediate conflict to settle is as destructive as any other unprovoked attack.

5 | Confusing Morality With Goodness

Moral judgment, then, is not a reliable guide to life in general. It is a useful guide for conflict but a terrible guide for anything else. If you devoted your whole life to morality and justice, your inevitable misjudgments would leave a trail of innocent victims in ruin. When all you have is the hammer of justice, everything calls for a sinner to be nailed to the cross.

But captivated by morality's powers, its admirers get carried away and take their praises to the grandest heights imaginable. They elevate morality to encompass all goodness, not only what is morally good but everything good. They call it the source of all values and the meaning of life. They credit morality with our friendly motives of benevolence, kindness, and generosity. Not to mention the story that morality can get your soul into heaven.

Such indiscriminate praise is deeply mistaken and hazardous. It encourages misjudgment by constantly calling for the moral sword. *We must fight for justice*, repeats the one-track moralist no matter the problem—*fight* instead of cooperate, understand, negotiate, or innovate. And why might hostilities rise when we preach that fighting is the essence of goodness itself? It helps zealots lure followers with themes of kindness and then convert them to hostile moralists to deploy against their enemies. It obscures the difference between morality and goodness so that we cannot distinguish when moral judgment is good or bad, productive or destructive, a defense or an attack. Under its spell, throwing stones at a blasphemer may seem as "good" as feeding the hungry.

So let us dispel this confusion by sharpening the distinctions between goodness, benevolence, and morality. We will use facts and ideas from evolutionary psychology, moral psychology, and natural language. We will take our time and first present some fundamentals of these subjects before we bring them together.

6 | Evolutionary Psychology

Evolutionary biology shows that humans are animals, a species of great ape who share a family with chimpanzees, bonobos, gorillas, and orangutans, as well as our extinct relatives like *Homo neanderthalensis*, *Homo erectus*, and *Australopithecus* [8, 9]. Humans were not created out of nothing by a supernatural designer. They evolved by natural selection. This may seem obvious, but it is the most important fact you need in order to understand anything about the human body and mind.

Natural selection created humans out of lifeless matter, starting billions of years ago with the first genetic replicators [8]. Once genes began to replicate and mutate, natural selection began to shape them, generation after generation, to make them better at competing for the limited resources they need for replication. Within the last 10 million years, natural selection forged humans out of the ape ancestors we share with chimpanzees and gorillas. Natural selection made humans human, chimpanzees chimpanzee, and gorillas gorilla by modifying the parts of their bodies and minds over hundreds of thousands of cycles of genetic replication, mutation, and competition for limited resources.

Much is known about natural selection, how it filters for genes that improve health and reproduction, and how it accumulates beneficial modifications to create complex adaptations made of countless parts within parts, such as the eye, the heart, and the hemoglobin molecule that carries oxygen in the blood of humans, birds, and fish alike. The science of natural selection provides us with a wealth of facts for our reasoning about human nature.

One of the fundamental insights is that an animal is essentially a bundle of adaptations, including cells, organs, appendages, and behaviors performed by programs stored in their brains. An animal's genes build these adaptations to protect and replicate themselves through reproduction. Each adaptation has been sculpted by natural selection to solve specific problems, such as hearts for pumping blood, legs for walking over land, and motor regions of the brain for operating the legs. As a result, adaptations evolve to be well-designed for a purpose, even though no planner designed them. Like tools and machines designed by humans, adaptations perform their functions by using principles of physics and engineering. Hearts are pumps that use hydrodynamics to circulate blood, eyes are sensors that use optics to form images, and so on. Therefore, to understand an animal is to understand its adaptations. A complete understanding would describe all of its adaptations, the problems they solve, the physical principles they use, and how their functions contribute to the animal's health and reproduction.

An animal's brain performs computations to perceive the world through its senses and to decide, plan, and direct the animal's actions. Brains mirror computers like hearts mirror pumps. Brains process information gathered from sensors, store data in memory, make decisions, plan actions, learn, and orchestrate behaviors. We can understand adaptations in the brain by using principles from computer science, decision theory, and game theory.

An animal's *mind* is all of the software—the programs and algorithms—that its brain executes [10, 11]. For example, the little brown bat has mental programs for echolocation, while migratory birds like the arctic tern have programs for migrating long distances, including algorithms that track the sun, stars, and Earth's magnetic field [12, 13]. Thus, the bat's mind can echolocate and perceive objects' shapes and positions, while the tern's mind can navigate by the sun and stars. These programs are physically encoded in the brain by genes, and their code evolves by natural selection just like cells, organs, and limbs. In fact, mental programs coevolve with the anatomy they operate. For example, an orb-weaver spider has seven types of glands for producing different types of silk, which coevolved with the web-spinning programs in its brain that know which silk to use for which part of the web [14]. The glands would be useless without the programs to operate them. What an animal perceives, knows, wants, feels, thinks, and remembers is part of its mind. To understand an animal's mind is to understand what programs it has, what problems they solve, what perceptions, knowledge, and plans they compute, and how they help the animal prosper.

Finally, animals have a special class of mental adaptations—strategies—for dealing with other animals who can think, plan, and learn in response to them [8]. In hunting, mating, fighting, and cooperation, an animal needs to plan actions toward someone who can act and plan in response. A strategy is a plan for dealing with a planner. Strategies are more complicated than regular plans because the problem you are trying to solve is another animal trying to solve the problem of you. Your strategy needs to consider the strategy of the other player, who will probably consider your strategy, and so on, layering strategies upon strategies with spiraling complexity.

The ways to solve these multiplayer problems are the subject of game theory. Thus, we need game theory to understand an animal's social behavior, just like we need hydrodynamics to understand the heart, and we need optics to understand the eye. An animal's mind evolves strategies such as hiding from predators or prey, confusing pursuers by dodging left or right unpredictably, cooperating with a mate to build a nest, assessing the strength of an opponent by the depth of their roar, and choosing a partner to hunt prey cooperatively. To understand these strategies, we analyze how they perform compared to other strategies, as we would study strategies in chess or poker. Animals play a lot of games, and their minds are full of evolved strategies.

We have now traced a path from natural selection through physical and mental adaptations to the evolved strategies of animals. We are, therefore, prepared to use these facts of evolution to understand the human mind. The human mind is an animal mind made of mental adaptations and strategies that help humans prosper in evolutionary games with other human players. We can understand our social minds by describing the games we play and the strategies humans evolved to play them. Particularly, we can examine social behaviors such as aggression, benevolence, hierarchy, and moral judgment as evolutionary strategies. We thereby anchor our reasoning in the realities of biology and physics, rather than drifting out to an endless sea of abstraction and confusion.

7 | Moral Psychology

Next, we call on moral psychology to supply us with facts about how moral judgment works. We want to understand moral judgment as a strategy in the mind that evolved for multiplayer games. We want to know which games the strategy is for and what other strategies it competes against.

To understand any adaptation, we consider its *form* and *function*, going back and forth between them to look for a close fit [9]. We hypothesize a function, use the hypothesis to make predictions about its form, observe facts about the form to test the hypothesis, and repeat these steps to find the best fit between a function and the observed form. When a key fits a lock, we have evidence that the key is designed for the lock. When the sphinx moth has an exceptionally long proboscis, up to 30 cm, that fits the exceptionally long nectar spur of the Madagascar star orchid, we have evidence that the function of the proboscis is to suck nectar from that orchid [15]. This is how we test theories about the functions of adaptations.

In contrast, we do not judge an adaptation's function only by its effects, because the effects could be a function or a byproduct of a different function. For example, the gazelle's thirst for water causes some gazelles to be eaten by crocodiles, but this effect is a byproduct of thirst, not its function. We can confirm this conclusion by looking at the form of thirst, such as how it is inhibited by the sight of a crocodile. Therefore, judging an adaptation's function is not the same as judging its effects. Function cannot be reduced to causation. We need to examine the form of an adaptation to judge its evolutionary functions.

Moral psychology shows us the form of moral judgment. In a typical study, participants judge the moral wrongness of different actions under different conditions. They may judge actions in isolation, such as abortion or infidelity, or actions taken by characters in real or hypothetical situations. Thousands of these studies describe the patterns in the form of moral judgment [16, 17]. They include surveys and experiments conducted across numerous societies with different cultures, languages, religions, and economies. While the content of moral rules differs strikingly across cultures, deeper patterns in moral judgment repeat across many cultures, such as the prominent role of actions and intentions. These are the facts we can use to determine its strategic functions. The basic patterns are as follows.

Moral judgment comes in two forms: conscience and condemnation [4]. Conscience judges one's own actions by moral rules, while condemnation judges other people's actions, particularly when they do not directly affect the condemner. Conscience and condemnation apply moral rules for different purposes that come from different strategic positions. Conscience is used by an actor to decide whether to take an action that is considered wrong and would affect a victim, typically before the actor has decided in order to influence their choice. Condemnation is used by an observer to judge an actor's action toward a victim, typically after the actor's action in order to decide whether to condemn them. The positions of actor and observer are as different as the positions of pitcher and batter in baseball. Confusing them is like assuming pitchers aim for home runs and batters aim for strikes. Yet this confusion is common in theories that assume conscience and condemnation have the same goals [4].

Of the two forms, condemnation is more distinctive and revealing. Conscience directs the person's own choices, like most mental programs. Condemnation, in contrast, judges other people's actions. Observers condemn actions that do not affect them and even the actions of strangers. This presents a mystery: What does the observer gain from condemnation? When condemnation motivates aggression, the mystery deepens: Condemners risk provoking retaliation, so what benefits make it worth the risk? As a distinctive form, condemnation offers a clue to the function of moral judgment. Discovering its purpose could also explain conscience as a defensive strategy to avoid actions that draw condemnation from observers.

Moral judgment focuses on a person's *action* and computes the wrongness of that action. People morally judge actions such as murder, theft, and lying, as well as idiosyncratic taboos such as worshipping foreign gods, eating taboo foods, and drawing forbidden pictures. We judge an action's wrongness in magnitudes ranging from perfectly right to slightly, moderately, and extremely

wrong. The greater the wrongness, the more we condemn the offender, tell others of the offense, feel anger and malice toward the offender, and call for their punishment.

Moral judgment's focus on actions is another distinctive form. You could easily overlook it because it is so familiar. But other judgments focus on the goals, the ends, the results rather than the actions, the means, the methods used to reach a goal. Consider, for example, judgments of prudence. Suppose a machine is about to cut off five of your fingers. The only way to save the five fingers is to push a finger from the other hand into the gears to stop the machine. Should you sacrifice one finger to save five? Yes, because the result, losing one finger, is better than losing five fingers. The means is terrible and painful, but it is worth it. Judgments of prudence focus on the results, and they assess actions by their results. As an observer, too, we would approve of someone who sacrificed one finger to save five.

Now, turn the situation into a moral dilemma and see how your judgment changes. The machine is about to cut off the pointer fingers of five people. The only way to stop the machine is to grab the worker next to you and shove their finger into the gears against their will (because your fingers are not long enough). Should you sacrifice one person's finger to save five people's fingers? Many people would say no, you cannot cut off one person's finger to save other people's fingers. At least, that is what research on similar dilemmas finds: People judge wrongs absolutely, rigidly, and categorically—not only by weighing the costs and benefits [4, 16, 17]. The action of cutting off someone's finger is morally wrong, no matter what result it aims to achieve. This special emphasis on actions differs markedly from judgments of prudence, value, efficiency, altruism, character, convention, and other matters.

Of course, some people do judge it acceptable to injure, kill, steal, or lie for a greater good. The result does have some sway, but the action's wrongness weighs against it and often overrides it, unlike in matters of prudence, where the result is paramount. In trolley dilemmas, for example, the percentage of people who condemn killing one person to save five people ranges from 90% to 10% across different varieties of the dilemma [17]. That is, holding constant the possible results—one death versus five deaths—people's moral judgments vary with the details of the action. Most people say you can flip a switch to turn the trolley toward one person to save five; in contrast, most people say you cannot push one person in front of the trolley to save five. In these and numerous other dilemmas, moral judgment focuses closely on the nature of the action: Pushing is worse than switching, actions are worse than omissions, killing someone as a means is worse than killing them as a side effect, and so on. Moral judgment carefully judges the nature of the action, including its category, causation, intentions, and knowledge, while showing less regard for the consequences.

The focus on actions also differs from our social judgments. Many of our social judgments focus on *persons* rather than *actions*. When we choose our friends, mates, and partners, we judge the person as a whole rather than their individual actions in isolation. When we form alliances and coalitions, we judge who is in the alliance and who is not. When we form hierarchies, we judge the ranks of different people. These and other social judgments focus on persons, differing from moral judgment's focus on actions.

Putting actions over persons is illustrated by the blindfold of Lady Justice, which symbolizes impartiality. Impartiality is a fundamental component of moral judgment [4–7]. We insist that people should morally judge actions impartially, blind to who the person is, and we suspiciously accuse others of flouting this ideal to favor their allies. Compare this notion to our other social judgments: Would you wear a blindfold to choose your friends, mates, or partners? To form hierarchies or alliances? It would be foolish or impossible. Hierarchies and alliances cannot be blind. Humans are not always impartial, of course, but we hold the ideal that moral judgment should be impartial. We do judge offenses impartially to some extent, sometimes admitting when our family, friends, and allies are in the wrong. And we claim our moral judgment is impartial, whether true or not, showing it is a desirable ideal worth faking. In contrast, no one claims they chose their spouse or friends without considering who they are.

The focus on actions also shapes the form of moral rules such as *thou shalt not kill* and *thou shalt not commit adultery*. Moral rules are among the distinctive products of moral judgment. In addition to judging people's actions in specific events, moral judgment also creates, learns, evaluates, and debates moral rules in general, apart from any specific event and in preparation for conflicts that could happen in the future.

Moral rules feature an action expressed with a verb, such as *kill*, *steal*, or *eat* (a forbidden food), which is preceded by a modal verb such as *can*, *may, must*, *shall*, or one of their negations such as *must not* [7]. This form shows the prominence of actions, the verbs, in moral laws. Meanwhile, the person in the rule is the generic *thou*, or an indefinite *whoever*, which refers to a person in general and no one in particular. This form echoes the blindfold by omitting a specific person while specifying the action that is forbidden, allowed, or required. Thus, right there in the form of moral laws we find prominent actions, a rich variety in countless rules, paired with a generic person who is no one in particular.

The action is the most variable component of moral rules. Humans moralize a vast array of verbs in laws, which declare that you *must not*, *must*, or *can* do particular actions, giving us prohibitions, duties, and rights. A community's set of rules comprises its moral code. The stunning variety of moral codes across societies reflects the different sets of verbs they moralize. This cultural variety comes from moral judgment's ability to moralize new actions and to debate whether to add them to the community's code. Indeed, humans are too good at making laws, which creates a new problem: a profusion of laws that inevitably contradict each other. Thus, they need to discuss and debate the rules to maintain a consistent moral code.

8 | The Method of Natural Language

The third idea is rather simple, and we have been practicing it all along. The idea is to use natural language to build our theories of morality while avoiding, whenever possible, artificial language such as jargon, contrived phrases, bureaucratic abbreviations, and made-up definitions [18–20].

Natural language is the third essential ingredient for understanding moral judgment. The first two won't help if you get lost in

the endless labyrinths of moral jargon. There is no escape from this overwhelming source of confusion without understanding the advantages of natural language. We have been looking for the simplest and most natural words for our subjects. If we instead recklessly used jargon such as *prosocial*, *distributive*, and *normative*, the theories themselves would be different and worse—worse for our reasoning, communication, and debate. Natural words are better for theories of morality, just as Arabic numerals are better than Roman numerals for mathematics.

The method is captured by George Orwell's rules for clear writing, especially: "Never use a long word where a short one will do," and "Never use a foreign phrase, a scientific word or a jargon word if you can think of an everyday English equivalent" (p. 264) [21].

It is also found in Fowler and Fowler's classic rules of vocabulary: "Prefer the familiar word to the far-fetched," "Prefer the concrete word to the abstract," "Prefer the single word to the circumlocution," and "Prefer the short word to the long" (1906) [22].

We take these maxims not only as prudent for writers but also as a method for building scientific theories. Call it the method of natural language. Theories are commonly formulated in language, and even mathematical and symbolic theories are interpreted and debated in language. Theories, therefore, are built out of words, and the choice of words determines their clarity, coherence, and accuracy. The maxims of clear writing double as maxims for building sound theories. They favor natural language over artificial jargon.

An obvious reason is that scientists need to discuss their theories in language, so the principles of reliable and efficient language can help scientists communicate, debate, and improve their theories. Another reason is that clear language helps us think clearly. It aids our reasoning in addition to communication. Indeed, Orwell extended its merits to thought itself:

> Modern English, especially written English, is full of bad habits which spread by imitation and which can be avoided if one is willing to take the necessary trouble. If one gets rid of these habits one can think more clearly (p. 253) [21].

The psychology of language reinforces and refines the maxims of clarity [18]. Humans evolved language for speech, so speech represents the most natural form of language. Writing is an invention that simulates speech, so effective writing mirrors clear speech. This means that spoken language sets the standard for natural language. The familiar words we use every day are generally the most reliable and efficient words. They are rigorously tested by daily use, where misunderstanding has immediate costs, unlike artificial jargon sheltered in academic journals.

Short words are efficient and fundamental. For efficiency, the human mind reserves the shortest words for our most fundamental concepts. When you need to refer to a concept many times a day, you can save your breath by making the word for it one or two syllables. For instance, among the 100 most frequent words in English, over 90% have one syllable, and the rest have two [23]. Words with one or two syllables continue to dominate the top 1000 words at over 85% and the top 5000 words at over 70%. Short words represent the essential machinery of thought.

In contrast, artificial jargon that is unfamiliar and long is less reliable and less efficient. Beyond the single word, scholars often combine words to make even longer names, such as *disadvantageous inequity aversion*, *negative strong reciprocity*, and *dispositional prosocial behavior*. When writers use one of these phrases repeatedly, it functions as a single word: an artificial compound with a heavy load of tongue-tying syllables. These phrases also illustrate an artificial grammar that is common in scholarly and bureaucratic language: piled modifiers [19]. In natural language, speakers rarely use more than one, perhaps two, modifiers before a noun, typically in the form of adjective noun, such as *pretty flower*. Scholars, however, pile adjectives and nouns before the head noun, such as *model-free reinforcement learning principles*, which is a noun adjective noun noun noun, and *group-based third-party punishment game*, a noun adjective adjective noun noun noun. This abuse of modifiers is artificial and grammatically obtuse. Natural language would use a richer grammar to specify the relations between nouns, such as *principles of learning by reinforcement without a model* and *a game of punishment with players from rival groups*, which use prepositions to articulate the relations between nouns instead of dumping them in a pile and leaving the reader to sort it out.

There are more forms of artificial language. Scholars overuse WEIRD abbreviations like MFT, DPT, MAC, VBN, and SVPM—to mention a few codewords from morality research. Some take it to absurd extremes, such as discussing TG, TPP, TPPG, and TPCG (varieties of trust and punishment games), or discussing the results purely in terms of H1a, H1b, H2a…H5c, assuming readers can perfectly recall a litany of overlapping codewords. They overuse prefixes and suffixes in obscure and made-up words, such as *heteronormativity*, *religitimation*, and *responsibilization*. Related, scholars habitually use verbs in noun forms converted with suffixes, such as *implementation* instead of the verb *implement*, often omitting the subject and object of the action. Finally, scholars make up their own definitions, coining peculiar meanings that disregard common usage. In contrast, the professionals who write dictionary definitions (lexicographers) study usage to describe what people mean by a word. They know that the meanings of words are conventions formed through practical use in communication, forged from many speakers' experiences of being understood and misunderstood. Scholars create confusion when they presume to define words at will, echoing Humpty Dumpty: "When *I* use a word, it means just what I choose it to mean—neither more nor less."

The method of natural language instead directs us to build theories by taking advantage of natural words and grammar. While every science can benefit, the sciences of society and morality especially stand to gain. Human behavior is not like atoms, viruses, or distant galaxies—removed from experience and lacking a vocabulary. Indeed, language evolved for communication about people and their actions. As writers say, a typical sentence tells us "who did what to whom." Eight of the 10 most frequent nouns are pronouns for people [23]. Many of the most

frequent verbs represent human speech (*say*, *mean*, *talk*), people's actions toward others (*give*, *take*, *help*), and people's minds, including perception, knowledge, and motives (*know*, *think*, *see*, *want*). They also include modal verbs (such as *can*, *cannot*, *must*), which we use to express logical reasoning as well as permissions, threats, promises, and moral rules. Natural language specializes in people and their strategies.

In particular, moral judgment is designed to be expressed in language because its strategic functions require us to discuss and debate our judgments with others. Natural language needs to be well-supplied with moral words to make morality as we know it possible at all. It therefore provides materials for our theories. We can understand morality in its own words.

Hence, we will favor natural words that are more frequent and more familiar. For example, we favor *good* over *positive*, *value* over *utility function*, and *help* over *facilitate*. These words cannot be replaced with artificial jargon without weakening the theory. We will mark with italics the first instance of each natural word used as an element in the theory.

That is enough about language. We can now combine the three strands—evolutionary psychology, moral psychology, and natural language—to draw the distinctions we seek: How does morality differ from benevolence and goodness?

9 | Goodness

Goodness is a fundamental quality of things that comes from our animal nature. Things that help an animal survive and reproduce—food, water, shelter, mates, and so on—are *good*. Things that hinder survival and reproduction—scarcity, injury, predators, pathogens, and so on—are *bad*.

In fact, we can apply good and bad, to some extent, to the circumstances of any living thing designed by natural selection, including animals, plants, bacteria, viruses, and the fundamental replicators: genes. These are players in the game of evolution, for whom things are good and bad. Once living things evolve to pursue goals, we can say the things that aid their pursuit are good for them, while things that block their pursuit are bad for them. In contrast, inanimate objects like rocks, mountains, planets, and stars do not replicate or reproduce, they have not evolved by natural selection over generations, and they do not have goals, so nothing can be good or bad for them.

I said to some extent because we are speaking partly metaphorically when we say that an animal or a gene is a player in an evolutionary game, that sunlight is good for a tree, or that a mutation is bad for a gene. Metaphors are powerful tools in science, but any metaphor can be overextended, so we should use them consciously rather than habitually as dead metaphors. Genes, for example, lie at the boundary between inanimate objects and living things. Like an ambiguous image that oscillates between different faces, genes can be seen as lifeless molecules, complex machines, or players in evolutionary competition. In reality, they are all of the above: part-molecule, part-machine, and part-player. Genes are extraordinary chimeras that defy our ordinary categories. But we can grasp their nature by taking each perspective in turn while seeking consistency among them.

Thus, the first signs of good and bad begin with the evolution of life. They arise simultaneously with genetic replication, needs for food and resources, self-propelled movement, and adaptations such as sensors, motors, and shells. Life is good and death is bad. Food is good and starvation is bad. Health is good and injury is bad. A strong shell is good and a broken shell is bad.

Goodness coincides with *value*. Goodness comes in amounts such as very good and fairly good, and the amount indicates a thing's value. The value of a thing is the amount of good or bad it does for a player. Values can also be called *costs* and *benefits*, invoking an economic metaphor. They can be added and subtracted to yield *profits* and *losses*, such as Darwin's reference to *profitable variations* in a species [24].

Goodness depends on the player. A thing is good *for* one player or another in the evolutionary game. Nothing is good without being good for a player or a group of players. What is good for one player may be bad for another. A deep crevice is good for a lizard and bad for a hawk who hunts lizards. But this does not mean that goodness is subjective or imaginary. The crevice is objectively good for the lizard and bad for the hawk. This is a matter of fact in the evolutionary game, determined by each animal's survival. Their values conflict, and the conflict is real.

From good and bad we need a few steps to derive *happiness* and *suffering*, and synonymously, *pleasure* and *pain*. Happiness and suffering refer to an animal's feelings, and feelings are adaptations in its mind designed to ultimately direct its actions. Happiness and pleasure are good feelings, while pain and suffering are bad feelings. Good feelings like pleasure mark a good thing, a thing that improves health and reproduction and is therefore worth pursuing, while bad feelings like pain mark a bad thing to avoid. Broadly, these feelings function as internal signals that alert multiple systems in the mind and body to an opportunity or danger, thereby coordinating the animal's response. For example, the pleasure of food signals an animal's memory to remember the actions used to get the food so they can be repeated in the future.

Unlike genes, an animal's brain may be said to hold an *idea* of goodness, represented by good and bad feelings. Genes are subject to good and bad, but they do not have an idea of good and bad because they do not have brains, feelings, or ideas. Genes are the mindless builders of minds rather than animals with minds of their own. Animals have ideas of good and bad, including feelings, values, motives, purposes, and concepts for categorizing good and bad things and for storing in memory good and bad experiences.

As feelings, pleasure and pain are not the same as good and bad, nor their foundation. Rather, they are feelings that measure, estimate, and guess what things will be good or bad for the animal's genes as players in the game of evolution. What is good for the genes is a matter of fact, not feelings, which will be determined ultimately by the animal's success in reproduction. The feelings make mistakes, for example, when an animal enjoys a food that poisons them. Moreover, many good and bad

TABLE 1 | Differences between goodness, benevolence, and morality.

	Goodness	Benevolence	Morality
Definition	A quality found in things that benefit a player in an evolutionary game	A player's strategy to do good for another player	A player's strategy for choosing sides in conflicts using impartial rules of action
Minimum players	One player	Two players	Four players: two fighters and two side-takers
Scope	All living things including animals and genes	Social animals including parents and cooperators	Human societies with conflicts inflamed by hierarchy and alliances
Related concepts	Life, value, health, benefits versus costs	Help, love, kindness, altruism, mutualism, cooperation, trade, welfare tradeoffs	Condemnation, conscience, *Thou shalt not* [action], moral laws, coordination game, moral code, moral debate, contradictory laws
Ideas in animal minds	Good and bad feelings, motives, purpose, goals, happiness versus suffering, pleasure versus pain, judgments of value	Sympathy, generosity, reputation, character, a good person	Right and wrong, moral rules of action, crime, guilt, proportionate punishment, impartiality, rule of law, justice
Origin	Four billion years ago with the origin of life. Ideas of goodness began 500 million years ago with the evolution of animal brains	Hundreds of millions of years ago with the evolution of social behavior in animals	Three hundred thousand years ago with the evolution of humans, or possibly *Homo* ancestors

things do not evoke any feelings because the animal does not need feelings to handle them. For instance, various processes of respiration, circulation, and digestion are good but operate without any feelings or awareness because the mind does not need to coordinate a response. Thus, what is good and bad for an animal is much broader than what feels good and bad.

Genes do not only create pleasure and pain but also mold them precisely to what is good for those genes in a given species and situation. For example, an animal is designed to endure pain when it benefits its genes. A mother bird endures injuries to defend her offspring against a fox. A male elephant seal endures the pain of battle to fight for a harem of females. Animals have many competing motives and feelings, all shaped in detail to benefit the genes that build those mental programs.

We have finished the story of good and bad. Table 1 summarizes the account of goodness and the features we will contrast with benevolence and morality. Goodness first appeared four billion years ago with the origin of life and genes. The idea of goodness, in the form of feelings, motives, purpose, and concepts, evolved in animal brains about 500 million years ago. Humans and morality play no essential role in the story. Humans share goodness with all living things, and we share ideas and feelings of goodness with thousands, perhaps millions, of animal species. Like other animals, the human sense of goodness has been tailored to our lifestyle, but the fundamentals remain the same. Food, shelter, survival, and reproduction are good—millions of animal species agree. Moral judgment influences what we see as good, but it does not, by any means, *create* our basic sense of goodness, value, or purpose. That is a preposterous confusion. You might as well believe that the human ability to make fire is the essence of goodness and the purpose of life. We will get to morality soon, but first we come to benevolence.

10 | Benevolence

So far, we have looked at goodness for a single player, whether an animal or a gene. We now add a second player to consider how animals do good and bad to each other.

An animal can take actions that aim to do good or bad to a second player. Call these moves a *good action* and a *bad action*. A good action *helps* the player, while a bad action *hurts* them. A good action is *benevolent*, while a bad action is *aggressive*. Giving food, water, or shelter to another player is benevolent. Attacking, injuring, or killing a player is aggressive.

Benevolence and aggression are adaptations in an animal's mind that direct them to help and hurt other animals. They are evolved *strategies* because they are adaptations for dealing with other players in multiplayer games. Natural selection favors benevolence and aggression when they benefit the genes that encode these strategies in the animal's brain. Genes themselves can also be benevolent and aggressive when they are designed to directly help and hurt other genes (as in intragenomic conflict [25]). As before, however, genes do not have feelings or ideas of benevolence, having no minds of their own.

Benevolence and aggression are widespread in nature, and they are central topics in evolutionary biology [8, 9]. Evolutionary biology further distinguishes whether the action is good or bad for the actor [26]. Within benevolence, an action that is good for the actor and good for the receiver is called mutualism, while an action that is bad for the actor and good for the receiver is called altruism or cooperation. Within aggression, an action that is good for the actor and bad for the receiver is called selfish, while an action that is bad for the actor and bad for the receiver is called spite.

Animals have evolved countless forms of benevolence. The most prominent form is altruism toward kin, especially parental care toward offspring. For example, the mammary glands, which give mammals their name, are designed to provide milk to offspring. They are essentially benevolence glands, organs designed to provide a good thing, milk, to another animal, the offspring. Moreover, the mammary glands coevolved with mental programs that direct their use, such as programs for nursing, for dividing milk among multiple offspring, and for weaning as the offspring mature. These mental programs—including motives, feelings, and concepts—are adaptations for benevolence because they are designed to help another player.

Outside of kinship, animals have evolved benevolent strategies for cooperating with mates, partners, and groups. Examples include cooperative parenting, cooperative hunting, sharing food, trading favors, and collective defense against predators, including vigilance, alarm calls, and mobbing [8, 27–32]. These benevolent strategies succeed in evolutionary competition when they benefit both the actor and receiver immediately, such as in cooperative hunting and defense, or in the long run, such as by trading favors. Conversely, unprofitable benevolence does not survive in evolutionary competition.

At the same time, cooperative relationships create opportunities for selfish strategies. For example, hyenas cooperate to take down big prey [32]. Once the kill is made, the large bounty opens the door for selfish strategies to grab a larger share of the meat. Thus, cooperative strategies commonly spur the evolution of selfish strategies, much like an abundance of prey spurs a rise in its predators.

Facing a mixture of cooperative and selfish players, an animal can benefit from telling the difference. Many animal species have evolved mental abilities to detect cheating and to keep track of who cheats and who cooperates [8]. Their minds summarize these observations as *reputations*. They use reputations to choose partners for cooperation and to avoid and punish cheaters. In turn, animals seek to create a good reputation and attract quality partners, so they display their benevolence and restrain and hide their selfishness. The ability to track reputations creates a competition for partners, favoring those who appear the most productive and benevolent [31]. Reputation thereby fosters benevolence by adding to its evolutionary benefits.

The evolution of reputation is the first sign of the idea of a *good person* and a *bad person*, as we say among humans. We also speak of good and bad character. In other species, it would be a good chimpanzee, a good raven, a good cleaner fish, and so on. More generally, we can say a good player and a bad player in the relevant evolutionary games, that is, games with benevolent and selfish actions that observers record in reputations. A good player does things that are good for you (you being the judge of character), and a bad player does things that are bad for you. A player who gives you resources and helps you is a good player. A player who takes resources and hurts you is a bad player.

When we go beyond two players to three or more, the accounting of good and bad gets complicated. If a player does good for multiple players and harms no one, then their action is purely benevolent, like with two players. But if a player's action aims to do good to some players and bad to other players, then it is a mixture of benevolence and aggression. Perhaps the most obvious example is conflict between coalitions. When a male dolphin supports his ally against an opponent, the dolphin aims to help one player and to hurt another player. The dolphin's support is good and cooperative from their ally's perspective, while it is bad and aggressive from the opponent's perspective. Thus, with three or more players, an animal's actions are often mixed, rather than purely benevolent or aggressive.

Altogether, benevolence is a common strategy in evolutionary games, and it is widespread in animals. Once an animal acts to help and hurt other animals, we have a good action and a bad action, benevolence and aggression. Once an animal can form reputations to track other players' actions, we have a good player and a bad player, and in humans, a good person and a bad person. We may observe that pure benevolence is, essentially by definition, the greatest good. Pure benevolence does good to everyone it affects and hurts no one. It is not only good to some and bad to others, not only good on balance or good on average. No one disagrees with pure benevolence, at least no one benevolent does. Those who aspire to the greatest good should pursue pure benevolence as much as possible.

As with goodness itself, moral judgment and moral rules play no essential role in the evolution of benevolence, including mutualism, altruism, cooperation, kindness, and friendliness. Nor do we need morality for the idea of reputation and good character. Humans certainly make moral rules *about* benevolent and aggressive actions, which contributes to the confusion, but that does not mean the rules are the primary source of those behaviors. Humans also make moral rules about what foods can be eaten, but it would be absurd to say that humans need morality to eat at all. We do not eat because we follow a moral law, *thou shalt eat*, nor do we help others and cooperate solely for moral reasons and under the threat of moralistic punishment. Countless animal species help each other and cooperate with no sign of moral judgment, moral laws, or a moral code that group members debate and amend. Human children less than a year old act benevolently toward others, such as retrieving a toy for someone who dropped it [33], before they can understand words, much less the particular moral code of their society. Morality is not the main source of benevolence. Let us see then where moral judgment enters the game of human society.

11 | Morality

To make our way to morality, let us pick up where we left off with aggression. We said that aggression is an animal's strategy to hurt another animal, to inflict costs, injury, or death, to do what is bad for them. But why do animals hurt each other?

The evolutionary game, remember, is a competition for limited resources. Every species and every organ and contrivance in nature has been sculpted in detail by generation after generation of ceaseless competition. If ever there is an abundance of food, the population multiplies until scarcity returns. Therefore, as players pursue food, shelter, mates, and so on, they inevitably run into opponents pursuing the same resources. Scarcity is the ultimate source of conflict.

Aggression is a strategy for winning scarce resources against other players. An animal tries to hurt an opponent to win the resource, either by disabling the opponent or by deterring them with threats and costs that compel them to retreat. In evolutionary biology, the simplest model of conflict is the hawk–dove game, in which two players want the same resource. They can choose a peaceful move, *dove*, or an aggressive move, *hawk*, which aims to injure the opponent if they persist.

In the hawk–dove game and other models of fighting, natural selection generally favors a mixture of aggression and peace [8]. When most players are peaceful, aggression does well by hurting others to win more resources. This is why animals hurt each other and why aggression is ubiquitous in nature. However, as aggression wins resources and becomes more common, aggressive players become more likely to meet each other and suffer the costs of injury. Thus, as aggression multiplies, it increasingly defeats itself in bloody battle. When aggression is too common to be profitable, then peaceful retreat becomes the better strategy. The result is a mixed equilibrium of aggression and peace. Thus, natural selection favors a mixture of both aggression and cautious retreat. This explains why animals frequently retreat from conflict and rarely pursue aggression to the extreme of all-out war and cannibalism. Indeed, naturalists have long marveled at the restrained rituals that govern fights in many animal species.

Since individuals differ in aggression and strength, animals benefit from assessing and remembering the toughness of other players [8, 34]. They summarize their observations in reputations for toughness, which track other players' abilities and motives to hurt them. They use these reputations to submissively defer to players who are tougher while challenging players who are weaker. This is the psychology of dominance and submission, which animals use to decide when fighting is profitable. In social animals, these reputations typically result in a dominance hierarchy in which the players converge on the same ranking of who dominates whom.

Some social animals have evolved a counterstrategy to dominance hierarchies: alliances [35]. By teaming up, weaker players combine their strength to dominate a stronger player. Humans take the alliance strategy far beyond other animals. Humans form alliances, counteralliances, and alliances of alliances in multiple levels, creating complex networks of loyalty [36]. As a result, human societies typically use alliances to suppress hierarchies [37]. Alliances, however, create new dangers. Now when players fight, allies join each side, and the injuries to everyone multiply.

The problem of alliances brings us to moral judgment. Moral judgment is a counterstrategy to alliances [5–7]. Rather than supporting their ally, the player sides against the fighter who has taken the most morally wrong action, according to the community's moral code. If enough players choose sides according to the moral code, they can prevent collisions between alliances that would deal injuries to everyone.

Moral judgment, then, is a strategy for choosing sides in conflicts. This function explains its distinctive form: the features described by moral psychology [5, 6]. Why does moral judgment direct us to condemn and hurt people? It is a strategy for fighting, and fighting means hurting people. The moral strategy uses moral rules to choose which side to attack. It aims to settle conflicts, but it is still an attack. Why do people condemn offenses that have not hurt them personally? With alliances, conflicts commonly escalate to recruit more and more people remote from the original fight. Observers can prevent the escalation by loudly condemning an offense before the conflict spreads to them, aiming to recruit everyone to oppose the wrongdoer before the group divides into factions.

Why does moral judgment focus on actions instead of goals or persons? The actions in moral rules function as signals that condemners use to coordinate aggression among observers with opposing loyalties to each side. The signals need to be observable and objective so that allies on both sides can agree on who is wrong. Goals are not a solid ground for agreement: They are hidden, and opposing goals are the source of conflict to begin with. Persons are no better: Rival alliances disagree on which person is better. Loyalty to persons is what causes alliances to clash in the first place. Why do people think moral judgment should be impartial? Again, only impartial judgments can provide a basis for agreement between rival alliances.

Why does moral judgment create and apply rules in forms like *thou shalt not kill*? Moral rules specify the actions that observers will use to coordinate and take the same side in conflicts. The rules pair a modal verb (such as *must not*, *must*, or *may*) with an action (such as *kill*, *steal*, or *lie*) to state which actions will signal the community to oppose and punish the offender. The person is left unspecified (*thou*, *whoever*) to make the rule impartial, applying equally across alliances and levels of hierarchy.

Why do humans make so many moral rules? Why do moral codes differ across cultures and over time? People fight over many things, and moral judgment is designed to make new rules for new conflicts. Moral judgment uses a formula to create rules by inserting any action into a template like *thou shalt not* [action]. We use the formula to create a multiplicity of rules, including destructive and contradictory rules. Then we need to debate and decide which rules will be in force. This process produces variation across cultures and over time as communities invent and choose different laws to govern the conflicts they currently encounter. Why do we disagree and fight over which moral laws should govern society? With many laws to choose from, each player uses their moral judgment to argue for the laws that are best for themselves.

Altogether, then, the form of moral judgment fits closely with the function of choosing sides in conflicts, as discussed in previous work (see Section 12). What does this function tell us about how morality relates to benevolence and goodness?

We found the function of moral judgment deep in the territory of aggression. Animals evolved to hurt their opponents, to dominate them in hierarchies, and, in some species, to combine forces in alliances to inflict more damage. Humans take alliances to the extreme and suffer costly battles as a result. Moral judgment evolved to settle these conflicts by directing aggression with moral rules to coordinate a majority against one side. If humans had no alliances in hostile standoffs, they could not have evolved moral judgment to govern them: no rules of action, no impartiality, no

thou shalt nots, no moral codes in limitless variety. Morality is forever bound to conflict.

Benevolence, in contrast, was not forged in conflict. Aggression is foreign to its gentle nature. Benevolence evolved to help someone, not to hurt someone. Its advantages come from kinship, mutual benefit, and trading favors, not from injury, attacks, and punishment. It is widespread across animal species, including the parental care and mammary glands that distinguish mammals. Birds, reptiles, amphibians, and fish all enjoy the bounties of mutualism, altruism, and cooperation. Benevolent strategies have been around for hundreds of millions of years, while the moral strategy with its strange codes of law is a special adaptation peculiar to humans. Benevolent judgments focus on the goal and the person, not impartial rules of action. Reputations for benevolence assess the good a person offers by their generosity and character, not to be confused with moral reputations, which assess how much a person abides by and upholds a particular moral code. Benevolence judges the ends, morality the means. Benevolence does good reliably, while morality gambles on punishing wrongdoers strategically.

Goodness is even more vast and ancient than benevolence and by no means beholden to morality. Goodness began with the evolution of life. Animals' ideas of goodness guide them to the resources they need to survive and reproduce. Humans possess a sense of goodness like every species in the animal kingdom. Human purpose, goals, and values do not depend on a moral code. We do not need a moral compass for direction because we have a goodness compass. If you were alone on a desert island, you would have no use for moral judgment or for benevolence. But you would still have purpose and values. We know that food, warmth, and health are good, while scarcity, cold, and disease are bad. We do not need moral laws to tell us so.

Now suppose a shipwreck brings dozens of people to the island. Suppose further that their minds have been altered to delete any concept of aggression. They cannot imagine or comprehend hurting another person, while everything else functions normally. With multiple players, benevolent strategies find their use, and everyone cooperates in abundance. Generosity is practiced, admired, and favored. But without aggression, there remains no use for morality, condemnation, or rules of action; no use for making rules, debating them, or assembling a moral code. Not until aggression resumes. Give them the ability to hurt and the motive of scarcity. Give them weapons and the strategy of alliances. Now they need a moral sword.

12 | Notes and Further Reading

I have gathered here notes and further reading. The present arguments elaborate on the theory that moral judgment evolved for choosing sides in conflicts. For a full account of the theory and a wide range of evidence from moral psychology, see DeScioli and Kurzban [5]. It is also summarized in DeScioli [6].

The side-taking theory stemmed from previous work by DeScioli and Kurzban [4], arguing that moral conscience is distinct from condemnation, condemnation is the primary function of moral judgment while conscience serves as defense, and the evolutionary function of moral judgment is not cooperation, as many scholars have presumed.

The side-taking theory was later applied to understand the evolutionary origin of laws [7]. This work distinguishes laws from threats and explains how the human ability to make limitless laws leads to a constant battle to control the laws.

On the fundamentals of evolutionary psychology, animal cognition, and evolutionary biology, see Tooby and Cosmides [10], Pinker [11], de Waal [12], and Dawkins [8]. For overviews of moral psychology, see Haidt [16] and Hauser [17]. For applications of evolutionary and moral psychology to politics, see Boyer [38], Petersen [39], and Weeden and Kurzban [40].

My argument for natural language builds on previous work about the faults of academic writing, particularly Pinker [18], DeScioli and Pinker [19], and more generally the cognitive psychology of language, such as Pinker [41]. The word frequencies that I mentioned come from the Corpus of Contemporary American English [23].

The difference between benevolence and morality mirrors Hume's distinction between benevolence and justice [42]. Benevolence, Hume observed, is doing good with a generous spirit, and it reflects "the highest merit, which *human nature* is capable of attaining" (p. 17). Justice, in contrast, serves to resolve conflicts with rules and laws. Without conflict, "the cautious, jealous virtue of justice would never once have been dreamed of" (p. 21). In a state of unlimited abundance, justice would serve no purpose: "Justice, in that case, being totally useless, would be an idle ceremonial, and could never possibly have place in the catalogue of virtues" (p. 21). I would not mind if someone said that the present arguments provide an evolutionary commentary on Hume's distinction.

Acknowledgments

I want to thank Natalie DeScioli, Robert Kurzban, and Andrew Delton for thoughtful comments, and special thanks to Natalie for suggesting the title.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. M. Brenan, *Views of State of Moral Values in U.S. at New Low*. (Gallup, 2023), https://news.gallup.com/poll/506960/views-state-moral-values-new-low.aspx.

2. A. M. Mastroianni and D. T. Gilbert, "The Illusion of Moral Decline," *Nature* 618, no. 7966 (2023): 782–789.

3. B. Russell, *Sceptical Essays* (Routledge, 1928/2004).

4. P. DeScioli and R. Kurzban, "Mysteries of Morality," *Cognition* 112 (2009): 281–299.

5. P. DeScioli and R. Kurzban, "A Solution to the Mysteries of Morality," *Psychological Bulletin* 139 (2013): 477–496.

6. P. DeScioli, "The Side-Taking Hypothesis for Moral Judgment," *Current Opinion in Psychology* 7 (2016): 23–27.

7. P. DeScioli, "On the Origin of Laws by Natural Selection," *Evolution and Human Behavior* 44 (2023): 195–209.

8. R. Dawkins, *The Selfish Gene, 40th Anniversary Edition* (Oxford University Press, 2016).

9. G. C. Williams, *The Pony Fish's Glow* (Basic Books, 1998).

10. J. Tooby and L. Cosmides "Psychological Foundations of Culture," in *The Adapted Mind*, ed. J. Barkow, L. Cosmides, and J. Tooby (Oxford University Press, 1992), 19–136.

11. S. Pinker, *How the Mind Works* (W. W. Norton & Company, 1997).

12. F. de Waal, *Are We Smart Enough to Know How Smart Animals Are?* (W. W. Norton & Company, 2016).

13. R. Wiltschko and W. Wiltschko, "Animal Navigation: How Animals Use Environmental Factors to Find Their Way," *European Physical Journal Special Topics* 232, no. 2 (2023): 237–252.

14. E. Pennisi, "Untangling Spider Biology," *Science* 358, no. 6361 (2017): 288–291.

15. J. Arditti, J. Elliott, I. J. Kitching, and L. T. Wasserthal, "Good Heavens What Insect Can Suck It'—Charles Darwin, *Angraecum sesquipedale* and *Xanthopan morganii praedicta*," *Botanical Journal of the Linnean Society* 169, no. 3 (2012): 403–432.

16. J. Haidt, *The Righteous Mind* (Vintage Books, 2012).

17. M. D. Hauser, *Moral Minds* (Springer, 2006).

18. S. Pinker, *The Sense of Style* (Penguin Books, 2014).

19. P. DeScioli and S. Pinker, "Piled Modifiers, Buried Verbs, and Other Turgid Prose in the American Political Science Review," *PS: Political Science and Politics* 55 (2022): 123–128.

20. P. DeScioli, "Computational Theories Should be Made With Natural Language Instead of Meaningless Code," *Behavioral and Brain Sciences* 46 (2023): e332.

21. G. Orwell, "Politics and the English Language," *Horizon* 13 (1946): 252–265.

22. H. W. Fowler and F. G. Fowler, *The King's English* (Clarendon Press, 1906).

23. M. Davies, "The Corpus of Contemporary American English," (2019), www.english-corpora.org/coca/.

24. C. Darwin, *On the Origin of Species* (John Murray, 1859).

25. A. Burt and R. Trivers, *Genes in Conflict* (Harvard University Press, 2006).

26. S. A. West, A. S. Griffin, and A. Gardner, "Social Semantics: Altruism, Cooperation, Mutualism, Strong Reciprocity and Group Selection," *Journal of Evolutionary Biology* 20, no. 2 (2007): 415–432.

27. J. Krause and G. D. Ruxton, *Living in Groups* (Oxford University Press, 2002).

28. D. Lukas and T. Clutton-Brock, "Cooperative Breeding and Monogamy in Mammalian Societies," *Proceedings of the Royal Society B: Biological Sciences* 279, no. 1736 (2012): 2151–2156.

29. S. W. Townsend, M. Rasmussen, T. Clutton-Brock, and M. B. Manser, "Flexible Alarm Calling in Meerkats: The Role of the Social Environment and Predation Urgency," *Behavioral Ecology* 23, no. 6 (2012): 1360–1364.

30. M. Dutour, J. P. Lena, and T. Lengagne, "Mobbing Behaviour Varies According to Predator Dangerousness and Occurrence," *Animal Behaviour* 119 (2016): 119–124.

31. R. Noë and P. Hammerstein, "Biological Markets," *Trends in Ecology & Evolution* 10, no. 8 (1995): 336–339.

32. K. E. Holekamp, S. T. Sakai, and B. L. Lundrigan, "Social Intelligence in the Spotted Hyena (*Crocuta crocuta*)," *Philosophical Transactions of the Royal Society B: Biological Sciences* 362 (2007): 523–538.

33. F. Warneken and M. Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees," *Science* 311, no. 5765 (2006): 1301–1303.

34. G. Arnott and R. W. Elwood, "Assessment of Fighting Ability in Animal Contests," *Animal Behaviour* 77, no. 5 (2009): 991–1004.

35. A. H. Harcourt, and F. de Waal, eds., *Coalitions and Alliances in Humans and Other Animals* (Oxford University Press, 1992).

36. P. DeScioli and E. O. Kimbrough, "Alliance Formation in a Side-Taking Experiment," *Journal of Experimental Political Science* 6 (2019): 53–70.

37. C. Boehm, *Hierarchy in the Forest* (Harvard University Press, 1999).

38. P. Boyer, *Minds Make Societies* (Yale University Press, 2018).

39. M. B. Petersen "Evolutionary Political Psychology," in *The Handbook of Evolutionary Psychology*, 2nd ed. (Wiley, 2015), 1084–1100.

40. J. Weeden and R. Kurzban, *The Hidden Agenda of the Political Mind* (Princeton University Press, 2014).

41. S. Pinker, *The Stuff of Thought* (Viking, 2007).

42. D. Hume, *An Enquiry Concerning the Principles of Morals* (A. Millar, 1751).