Game Theory and Morality

Moshe Hoffman, Erez Yoeli, and Carlos David Navarrete

Introduction

Consider the following puzzling aspects of our morality:

1. Many of us share the view that one should not use people, even if it benefits them to be used, as Kant intoned in his second formulation of the categorical imperative: "Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end" (Kant, 1997). Consider dwarf tossing, where dwarfs wearing protective padding are thrown for amusement, usually at a party or pub. It is viewed as a violation of dwarfs' basic dignity to use them as a means for amusement, even though dwarves willingly engage in the activity for economic gain. Many jurisdictions ban dwarf tossing on the grounds that the activity violates dwarfs' basic human rights, and these laws have withstood lawsuits raised by dwarfs suing over the loss of employment (!).
2. Charitable giving is considered virtuous, but little attention is paid to how just the cause or efficient the charity. For example, Jewish and Christian traditions advocate giving 10 % of one's income to charity, but make no mention of the importance of evaluating the cause or avoiding wasteful charities. The intuition that giving to charity is a moral good regardless of efficacy results in the persistence of numerous inefficient and corrupt charities. For example, the Wishing Well Foundation has, for nearly a decade, ranked as one of CharityNavigator.

M. Hoffman (✉) • E. Yoeli
Program for Evolutionary Dynamics, Harvard University,
One Brattle Square, Suite 6, Cambridge, MA 02138, USA
e-mail: moshehoffman@fas.harvard.edu

C.D. Navarrete
Department of Psychology, and, the Ecology, Evolutionary Biology and Behavior Program,
Michigan State University, East Lansing, MI, USA

T.K. Shackelford, R.D. Hansen (eds.), *The Evolution of Morality*,
Evolutionary Psychology, DOI 10.1007/978-3-319-19671-8_14

com's most inefficient charities. Yet its mission of fulfilling wishes by children with terminal illnesses is identical to that of the more efficient Make-A-Wish Foundation. Worse yet, scams masquerading as charities persist. One man operating as The US Navy Veteran's Association collected over 100 million dollars—over 7 years!—before anyone bothered to investigate the charity.

3. In every culture and age, injunctions against murder have existed. If there is one thing much of humanity seems to agree on, it's that ending the life of another without just cause which is among the worst of moral violations. Yet cultures don't consider the loss of useful life years in their definition, even though it is relevant to the measure of harm done by the murder. Why is our morality so much more sensitive to *whether* a life was lost than to how much life was lost?

There are numerous other examples of how our moral intuitions appear to be rife with logical inconsistencies. In this chapter, we use game theory to provide insight on a range of moral puzzles similar to the puzzles described above.

What Is Game Theory and Why Is It Relevant?

In this section, we review the definition of a game, and of a Nash equilibrium, then discuss how evolution and learning processes would yield moral intuitions consistent with Nash equilibria.

Game theory is a tool for the analysis of social interactions. In a game, the *payoff* to each *player* depends on their actions, as well as the actions of others. Consider the Prisoner's Dilemma (Chammah & Rapoport, 1965; see Fig. 1), a model that captures the paradox of cooperation. Each of two players chooses whether to cooperate or to defect. Cooperating reduces a player's payoff by $c > 0$ while increasing the other's payoffs by $b > c$. Players could be vampire bats with the option of sharing blood, or firms with the option of letting each other use their databases, or premed students deciding whether to take the time to help one another to study. The payoffs, b and c, may represent likelihood of surviving and leaving offspring, profits, or chance of getting into a good medical school.

Solutions to such games are analyzed using the concept of a Nash equilibrium[1]—a specification of each player's action such that no player can increase his payoff by deviating unilaterally. In the Prisoner's Dilemma, the only Nash equilibrium is for neither player to cooperate, since regardless of what the other player does, cooperation reduces one's own payoff.

[1] Note that we focus on the concept of Nash equilibrium in this chapter and not evolutionary stable strategy (ESS), a refinement of Nash that might be more familiar to an evolutionary audience. ESS are the Nash equilibria that are most relevant in evolutionary contexts. However, ESS is not well defined in many of our games, so we will focus on the insights garnered from Nash and directly discuss evolutionary dynamics when appropriate.

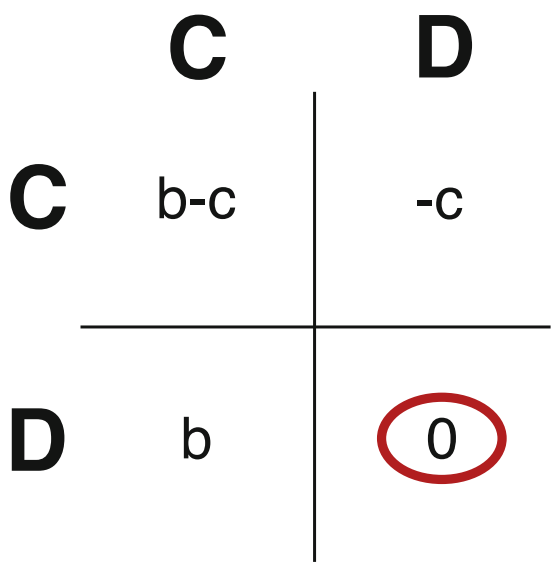

Fig. 1 The Prisoner's Dilemma. Player 1's available strategies (C and D, which stand for cooperate and defect, respectively) are represented as *rows*. Player 2's available strategies (also C and D) are represented as *columns*. Player 1's payoffs are represented at the intersection of each row and column. For example, if player 1 plays D and player 2 plays C, player 1's payoff is b. The Nash equilibrium of the game is (D, D). It is indicated with a *circle*

Game theory has traditionally been applied in situations where players are rational decision makers who deliberately maximize their payoffs, such as pricing decisions of firms (Tirole, 1988) or bidding in auctions (Milgrom & Weber, 1982). In these contexts, behavior is expected to be consistent with a Nash equilibrium, otherwise one of the agents—who are actively deliberating about what to do—would realize she could benefit from deviating from the prescribed strategy.

However, game theory also applies to evolutionary and learning processes, where agents do not deliberately choose their behavior in the game, but play according to strategies with which they are born, imitate, or otherwise learn. Agents play a game and then "reproduce" based on their payoffs, where reproduction represents offspring, imitation, or learning. The new generation then play the game, and so on. In such settings, if a mutant does better (mutation can be genetic or can happen when agents experiment), then she is more likely to reproduce or her behavior imitated or reinforced, causing the behavior to spread. This intuition is formalized using models of evolutionary dynamics (e.g., Nowak, 2006).

The key result for evolutionary dynamic models is that, except under extreme conditions, behavior converges to Nash equilibria. This result rests on one simple, noncontroversial assumption shared by all evolutionary dynamics: Behaviors that are relatively successful will increase in frequency. Based on this logic, game theory models have been fruitfully applied in biological contexts to explain phenomena such as animal sex ratios (Fisher, 1958), territoriality (Smith & Price, 1973), cooperation (Trivers, 1971), sexual displays (Zahavi, 1975), and parent–offspring conflict (Trivers, 1974). More recently, evolutionary dynamic models have been applied in human contexts where conscious deliberation is believed to not play an important role, such as in the adoption of religious rituals (Sosis & Alcorta, 2003), in the expression and experience of emotion (Frank, 1988; Winter, 2014), and in the use of indirect speech (Pinker, Nowak, & Lee, 2008).

Crucially for this chapter, because our behaviors are mediated by moral intuitions and ideologies, if our moral behaviors converge to Nash, so must the intuitions and ideologies that motivate them. The resulting intuitions and ideologies will bear the signature of their game theoretic origins, and this signature will lend clarity on the puzzling, counterintuitive, and otherwise hard-to-explain features of our moral intuitions, as exemplified by our motivating examples.

In order for game theory to be relevant to understanding our moral intuitions and ideologies, we need only the following simple assumption: *Moral intuitions and ideologies that lead to higher payoffs become more frequent*. This assumption can be met if moral intuitions that yield higher payoffs are held more tenaciously, are more likely to be imitated, or are genetically encoded. For example, if every time you transgress by commission you are punished, but every time you transgress by omission you are not, you will start to intuit that commission is worse than omission.

Rights and the Hawk–Dove Game

In this section we will argue that just as the Hawk–Dove model explains animal territoriality (Maynard Smith & Price, 1973, *to be reviewed shortly*), *the Hawk–Dove model sheds light onto our sense of rights* (Descioli & Karpoff, 2014; Gintis, 2007; Myerson, 2004).

Let us begin by asking the following question (Myerson, 2004): "Why [does] a passenger pay a taxi driver after getting out of the cab in a city where she is visiting for one day, not expecting to return?" If the cabby complains to the authorities, the passenger could plausibly claim that she had paid in cash. The answer, of course, is that the cabby would feel that the money the passenger withheld was his—that he had a right to be paid for his service—and get angry, perhaps making a scene or even starting a fight. Likewise, if the passenger did in fact pay, but the cabby demanded money a second time, the passenger would similarly be infuriated. This example illustrates that people have powerful intuitions regarding rightful ownership. In this section, we explore what the Hawk–Dove game can teach us about our sense of property rights.

The reader is likely familiar with the Hawk–Dove game, a model of disputes over contested resources. In the Hawk–Dove game, each player decides whether to fight over a resource or to acquiesce (i.e. play Hawk or Dove). If one fights and the other does not, the fighter gets the resource, worth v. If both fight, each pays a cost c and split the resource. That is, each gets $v/2-c$. If neither fights, they split the resource and get $v/2$. As long as $v/2 < c$, then in any stable Nash equilibrium, one player fights and the other acquiesces. That is, if one player expects the other to fight, she is better off acquiescing, and vice versa (see Fig. 2).

Crucially, it is not just a Nash equilibrium for one player to always play Hawk and the other to always play Dove. It is also an equilibrium for both players to condition whether they play Hawk on an *uncorrelated asymmetry*—a cue or event that

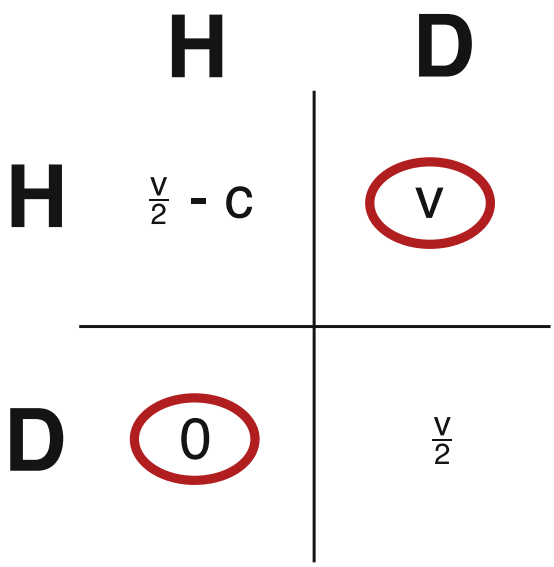

Fig. 2 The Hawk–Dove game. The Nash equilibria of the game are circled

does not necessarily affect the payoffs, but does distinguish between the players, such as who arrived at the territory first or who built the object. If one conditions on the event (say, plays Hawk when she arrives first), then it is optimal for the other to condition on the event (to play Dove when the other arrives first).

As our reader is likely aware, this was the logic provided by Maynard Smith to explain animal territoriality—why animals behave aggressively to defend territory that they have arrived at first, even if incumbency does not provide a defensive advantage and even when facing a more formidable intruder. Over the years, evidence has amassed to support Maynard Smith's explanation, such as experimental manipulation of which animal arrives first (Davies, 1978; Sigg & Falett, 1985).

Like other animals, we condition how aggressively we defend a resource on whether we arrive first. Because our behaviors are motivated by beliefs, we are also more likely to believe that the resource is "ours" when we arrive first. Studies have shown these effects with children's judgments of ownership, in ethnographies of prelegal societies, and in computer games. In one such illustration, DeScioli and Wilson (2011) had research subjects play a computer game in which they contested a berry patch. Subjects who ended up keeping control of the patch usually arrived first, and this determined the outcome more often than differences in fighting ability in the game.

This sense of ownership is codified in our legal systems, as illustrated by the quip "possession is 9/10ths of the law," and in a study involving famous legal property cases conducted by Descioli and Karpoff (2014). In a survey, these researchers asked participants to identify the rightful owner of a lost item, after reading vignettes based on famous property rights legal cases. Participants consistently identified the possessor of the found item as its rightful owner (as the judges had at the time of the case). This sense of ownership is also codified in our philosophical tradition, e.g., in Locke (1988), who found property rights in initial possession. Note that, as has also been found in animals, possession extends to objects on one's land: In DeScioli and

Karpoff's survey, another dictate of participants' (and the judges') property rights intuitions was who owned the land on which the lost item was found.

Also like animals, our sense of property rights is influenced by who created or invested in the resource, another uncorrelated asymmetry. In locales that sometimes grant property rights to squatters—individuals who occupy lands others have purchased—a key determinant of whether the squatters are granted the land is whether they have invested in it (Cone vs. West Virginia Pulp & Paper Co., 1947; Neuwirth, 2005). Locke also intuited that investment in land is part of what makes it ours: In *Second Treatise on Civil Government* (1689), Locke wrote, "everyman has a property in his person; this nobody has a right to but himself. The labor of his body and the work of his hand, we may say, are properly his."

If the Hawk–Dove model underlies our sense of property rights, we would expect to see psychological mechanisms that motivate us to feel entitled to an object when we possess it or have invested in it. Here are three such mechanisms, which can be seen by reinterpreting some well-documented "biases" in the behavioral economics literature. The first such bias is the *endowment effect*: We value items more if we are in possession of them. The endowment effect has been documented in dozens of experiments, where subjects are randomly given an item (mug, pen, etc.) and subsequently state that they are willing to sell the mug for much more than those who were not given the mug are willing to pay (Kahneman, Knetsch, & Thaler, 1990). In the behavioral economics literature, the endowment effect has sometimes been explained by loss aversion, which is when we are harmed more by a loss than we benefit from an equivalent gain. However, the source of loss aversion is not questioned or explained. When it is, loss aversion is also readily explained by the Hawk–Dove game (Gintis, 2007).

A second bias that also fits the Hawk–Dove model is the *IKEA effect*: Our valuation of an object is influenced by whether we have developed or built the resource. The IKEA effect has been documented by asking people how much they would pay for items like Lego structures or IKEA furniture after randomly being assigned to build them or receive them pre-built. Subjects are willing to pay more for items they build themselves.

A third such bias that fits the Hawk–Dove model is the *sunk cost fallacy* (Mankiw, 2007; Thaler, 1980), which leads us to "throw good money after bad" when we invest in ventures simply because we have already put so much effort into them, arguably because our prior efforts lead us to value those ventures more.

Possession and past investment are not the only uncorrelated asymmetries that can dictate rights. Rights can be dictated by a history of agreements, as happens when one party sells another deed to a house or car, or, as in our taxicab example, by whether a service was provided. There are also countless examples in which rights were determined by perhaps unfair or arbitrary characteristics such as race and sex: Black Americans were expected to give up their seat for Whites in the Jim Crow South and women to hand over their earnings or property to their husbands throughout the ages.

Hawk–Dove is not just a post hoc explanation for our sense of rights; it also leads to the following novel insight: We can formally characterize the properties that

uncorrelated asymmetries must have. This requires a bit more game theory to illustrate; the logic is detailed in the section on categorical distinctions but the implications are straightforward: Uncorrelated asymmetries must be discrete (as in who arrived first or whether someone has African ancestry) and cannot be continuous (who is stronger, whether someone has darker skin). Indeed, we challenge the reader to identify a case where our sense of rights depends on surpassing a threshold in a continuous variable (stronger than? darker than?). More generally, an asymmetry must have the characteristic that, when it occurs, every observer believes it occurred with a sufficiently high probability, where the exact level of confidence is determined by the payoffs of the game. This is true of public, explicit speech and handshakes, but not innuendos or rumors. (Formally, explicit speech and handshakes induce what game theorists term common *p*-beliefs.)

The Hawk–Dove explanation of our sense of rights also gives useful clarity on when there will be conflict. Conflict will arise if both players receive opposing signals regarding the uncorrelated asymmetry, such as two individuals each believing they arrived first, or when there are two uncorrelated asymmetries that point in conflicting directions, such as when one person invested more and the other arrived first. The former source of conflict appears to be the case in the Israeli–Palestinian conflict. Indeed, both sides pour great resources into demonstrating their early possession, especially Israel, through investments in and public displays of archeology and history. The latter source of conflict appears to be the case in many of the contested legal disputes in the study by DeScioli and Karpoff (2014) mentioned above. An example is one person finds an object on another's land. Indeed, this turns out to be a source of many legal conflicts over property rights, and a rich legal tradition has developed to assign precedence to one uncorrelated asymmetry over another (Descioli & Karpoff, 2014). As usual, we see similar behavior in animals in studies that provide empirical support for Maynard Smith's model for animal territoriality: When two animals are each given the impression they arrived first by, for example, clever use of mirrors, a fight ensues (Davies, 1978).

Authentic Altruism, Motives, and the Envelope Game

In this section, we present a simple extension of the Repeated Prisoner's Dilemma to explain why morality depends not just on what people do but also what they think or consider.

In the Repeated Prisoner's Dilemma and other models of cooperation, players judge others by their actions—whether they cooperate or defect. However, we not only care about whether others cooperate but also about their decision-making process: We place more trust in cooperators who never even considered defecting. To quote Kant, "In law a man is guilty when he violates the rights of others. In ethics he is guilty if he only thinks of doing so."

The Envelope Game (Fig. 3) models why we care about thoughts and considerations and not just actions (Hoffman, Yoeli, & Nowak, 2015). The Envelope Game

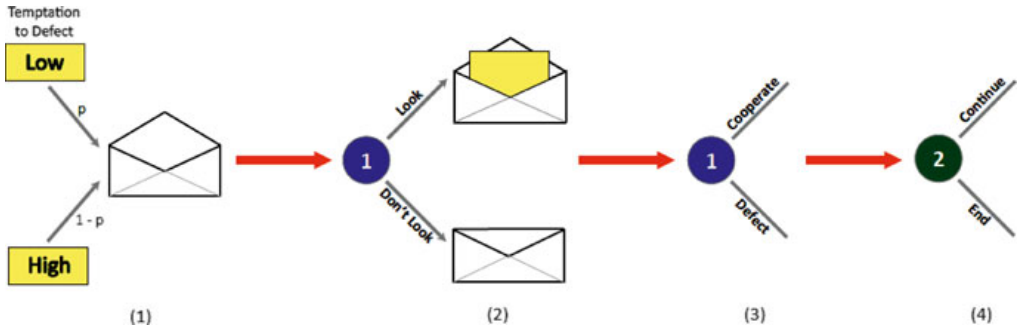

Fig. 3 A single stage of the Envelope Game

is a repeated game with two players. In each round, player 1 receives a sealed envelope, which contains a card stating the costs of cooperation (high temptation to defect vs. low temptation to defect). The temptation is assigned randomly and is usually low. Player 1 can choose to look inside the envelope and thus find out the magnitude of the temptation or choose not to look. Then player 1 decides to cooperate or to defect. Subsequently, player 2 can either continue to the next round or end the game. As in the Repeated Prisoner's Dilemma, the interaction repeats with a given likelihood, and if it does, an envelope is stuffed with a new card and presented to player 1, etc.

In this model, as long as temptations are rare, large, and harmful to player 2, it is a Nash equilibrium for player 1 to "cooperate without looking" in the envelope and for player 2 to continue if and only if player 1 has cooperated and not looked. We refer to this as the *cooperate without looking* (CWOL) equilibrium.[2] This equilibrium emerges in agent-based simulations of evolution and learning processes.[3] Notice that if player 1 could not avoid looking inside the envelope, or player 2 could not observe whether player 1 looked, there would not be a cooperative equilibrium since player 1 would benefit by deviating to defection in the face of large temptations. Not looking permits cooperative equilibria in the face of large temptations.

The Envelope Game is meant to capture the essential features of many interesting aspects of our morality, as described next.

Authentic Altruism. Many have asked whether "[doing good is] always and exclusively motivated by the prospect of some benefit for ourselves, however subtle" (Batson, 2014), for example, the conscious anticipation of feeling good (Andreoni,

[2] Technically, the conditions under which we expect players to avoid looking and attend to looking are $c_h > a/(1-w) > c_l p + c_h(1-p)$ and $bp + d(1-p) < 0$), where c_h and c_l are the magnitudes of the high and low temptations, respectively; p is the likelihood of the low temptation; $a/(1-w)$ is the value of a repeated, cooperative interaction to player 1; and $bp + d(1-p)$ is the expected payoff to player 2 if player 1 only cooperates when the temptation is low.

[3] The simulations employ numerical estimation of the replicator dynamics for a limited strategy space: cooperate without looking, cooperate with looking, look and cooperate only when the temptation is low, and always defect for player 1, and end if player 1 looks, end if player 1 defects, and always end for player 2.

1990), avoidance of guilt (Cain, Dana, & Newman, 2014; Dana, Cain, & Dawes, 2006; DellaVigna, List, & Malmendier, 2012), anticipation of reputational benefits or reciprocity (as Plato's Glaucon suggests, when he proffers that even a pious man would do evil if given a ring that makes him invisible; Trivers, 1971). At the extreme, this amounts to asking if saintly individuals such as Gandhi or Mother Teresa were motivated thus, or if they were "authentic" altruists who did good without anticipating any reward and would be altruistic even in the absence of such rewards. Certainly, religions advocate doing good for the "right" reasons. In the Gospel of Matthew, Chapter 6, Jesus advocates, "Be careful not to practice your righteousness in front of others to be seen by them. If you do, you will have no reward from your Father in heaven," after which he adds, "But when you give to the needy, do not let your left hand know what your right hand is doing, so that your giving may be in secret. Then your Father, who sees what is done in secret, will reward you."

The Envelope Game suggests authentic altruism is indeed possible: By focusing entirely on the benefits to others and ignoring the benefits to themselves, authentic altruists are trusted more, and the benefits from this trust outweigh the risk of, for example, dying a martyr's death. Moreover, this model helps explain why we think so highly of authentic altruists, as compared to others who do good, but with an ulterior motive (consider, as an example, the mockery Sean Penn has faced for showing up at disaster sites such as Haiti and Katrina with a photographer in tow).

Principles. Why do we like people who are "principled" and not those who are "strategic"? For example, we trust candidates for political office whose policies are the result of their convictions and are consistent over time and distrust those whose policies are carefully constructed in consultation with their pollsters and who "flip-flop" in response to public opinion (as caricatured by the infamous 2004 Republican presidential campaign television ad showing John Kerry windsurfing and tacking from one direction to another). CWOL offers the following potential explanation. Someone who is strategic considers the costs and benefits to themselves of every decision and will defect when faced with a large temptation, whereas someone who is guided by principles is less sensitive to the costs and benefits are to themselves and thus less likely to defect. Imagine our flip-flopping politician was once against gay marriage but supports it now that it is popular. This indicates the politician is unlikely to fight for the cause if it later becomes unpopular with constituents or risks losing a big donor. Moreover, this model may help explain why ideologues that are wholly devoted to a cause (e.g., Hitler, Martin Luther King, and Gandhi) are able to attract so many followers.

Don't Use People. Recall Kant's second formulation of the categorical imperative: "Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means but always at the same time as an end." In thinking this through, let's again consider dwarf tossing. Many see it as a violation of dwarfs' basic dignity to use them as a means for amusement, even though they willingly engage in the activity for economic gain. Our aversion to using people may explain many important aspects of our moral intuitions, such as

why we judge torture as worse than imprisonment or punishment (torture is harming someone as a means to obtaining information) and perhaps one of the (many) reasons we oppose prostitution (prostitution is having sex with someone as a means to obtaining money). The Envelope Game clarifies the function of adhering to this maxim. Whereas those who treat someone well as means to an end would also mistreat them if expedient, those who treat someone well as an end can be trusted not to mistreat them when expedient.

Attention to Motives. The previous two applications are examples of a more general phenomenon: that we judge the moral worth of an action based on the motivation of the actor, as argued by deontological ethicists, but contested by consequentialists. The deontological argument is famously invoked by Kant: "Action from duty has its moral worth not in the purpose to be attained by it but in the maxim in accordance with which it is decided upon, and therefore does not depend upon the realization of the object of the action but merely upon the principle of volition in accordance with which the action is done without regard for any object of the faculty of desire" (Kant, 1997). These applications illustrate that we attend to motives because they provide valuable information on whether the actor can be trusted to treat others well even when it is not in her interest.

Altruism Without Prospect of Reciprocation. CWOL also helps explain why people cooperate in contexts where there is no possibility of reciprocation, such as in one-shot anonymous laboratory experiments like the dictator game (Fehr & Fischbacher, 2003), as well as when performing heroic and dangerous acts. Consider soldiers who throw themselves on a grenade to save their compatriots or stories like that of Liviu Librescu, a professor at the University of Virginia and a Holocaust survivor, who saved his students during a school shooting. When he heard the shooter coming toward his classroom, Librescu stood behind the door to his classroom, expecting that when the shooter tried to shoot through the door, it would kill him and his dead body would block the door. Mr. Librescu, clearly, did not expect this act to be reciprocated. Such examples have been used as evidence for group selection (Wilson, 2006), but can be explained by individuals "not looking" at the chance of future reciprocation. Consistent with this interpretation, cooperation during extreme acts of altruism is more likely to be intuitive than deliberative (Rand & Epstein, 2014), and those who cooperate without considering the prospect of reciprocation are more trusted (Critcher, Inbar, & Pizarro, 2013). We also predict that people are more likely to cooperate intuitively when they know they are being observed.

The Omission–Commission Distinction and Higher-Order Beliefs

We explain the omission–commission distinction and the means–by-product distinction by arguing that these moral intuitions evolved in contexts where punishment is coordinated. Then, even when intentions are clear to one witness for omissions and by-products, a witness will think intentions are less clear to the other witnesses.

Why don't we consider it murder to let someone die that we could have easily saved? For example, we sometimes treat ourselves to a nice meal at a fancy restaurant rather than donating the cost of that meal to a charity that fights deadly diseases. This extreme example illustrates a general phenomenon: that people have a tendency to assess harmful commissions (actions such as killing someone) as worse, or more morally reprehensible, than equally harmful omissions (inactions such as letting someone die). Examples of this distinction abound, in ethics (we assess withholding the truth as less wrong than lying (Spranca, Minsk, & Baron, 1991)), in law (it is legal to turn off a patient's life support and let the patient die, as long as one has the consent of the patient's family; however, it is illegal to assist the patient in committing suicide even with the family's consent), and in international relations. For example, consider the Struma, a boat carrying Jewish refugees fleeing Nazi persecution in 1942. En route to Palestine, the ship's engine failed, and it was towed to a nearby port in Turkey. At the behest of the British authorities then in control of Palestine, passengers were denied permission to disembark and find their way to Palestine by land. For weeks, the ship sat at port. Passengers were brought only minimal supplies, and their requests for safe haven were repeatedly denied by the British and others. Finally, the ship was towed to known hostile waters in the Black Sea, where it was torpedoed by a Russian submarine almost immediately, killing 791 of 792 passengers. Crucially, though, the British did not torpedo the ship themselves or otherwise execute passengers—an act of commission that they and their superiors would undoubtedly have found morally reprehensible.

Why do we distinguish between transgressions of omission and commission? To address this question, we present a simple game theory model based on the insight by DeScioli, Bruening, and Kurzban (2011). The intuition can be summarized in four steps:

1. We note that moral condemnation motivates us to punish transgressors. Such punishment is potentially costly, e.g., due to the risk of retaliation. We expect people to learn or evolve to morally condemn only when such costs are worth paying.
2. Moral condemnation can be less costly when others also condemn, perhaps because the risk of retaliation is diffused, because some sanctions do not work unless universally enforced or, worse, because others may sanction individuals they believe wrongly sanctioned. This can be modeled using any game with multiple Nash equilibria, including the Repeated Prisoner's Dilemma and the Side-Taking Game. The Coordination Game is the simplest game with multiple equilibria, so we present this game to convey the basic intuition. In the Coordination Game, there are two players who each simultaneously choose between two actions, say punish and don't punish. The key assumption is that each player prefers to do what she expects the other to do, which can be captured by assuming each receives a if they both punish, d if neither punish, $b<d$ if one punishes and the other does not, and $c<a$ if one does not punish while the other does (Fig. 4).
3. Transgressions of omission that are intended are difficult to distinguish from unintended transgression, as is the case when perpetrators are simply not paying

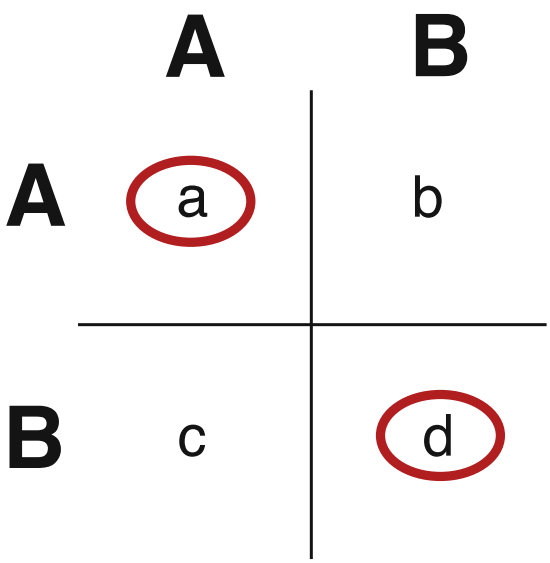

Fig. 4 The Coordination Game. In our applications, A stands for punish, and B stands for don't punish

attention or do not have enough time to react with better judgment (DeScioli et al., 2011). Relative to the example of the tennis player with the allergy described above, it is usually hard to distinguish between a competitor who does not notice his opponent orders the dish with the allergen versus one who notices but does not care. In contrast, transgressions of commission must be intended almost by definition.

4. Suppose the witness knows an omission was intentional: In the above example, the tennis player's opponent's allergy is widely known, and the witness saw the player watch his opponent order the offending dish, had time to react, thought about it, but did not to say anything. The witness suspects that others do not know the competitor was aware his opponent ordered the dish, but believes the tennis player should be condemned for purposely withholding information from his competitor. However, since the witness does not wish to be the sole condemner, she is unlikely to condemn. In contrast, when a witness observes a transgression of commission (e.g., the player recommends the dish), the witness is relatively confident that others present interpret the transgression as purposely harmful, since his recommendation reveals that the player was obviously paying attention and therefore intended to harm his opponent. So, if all other individuals present condemn the tennis player when they observe the commission, each does not anticipate being the sole condemner.

For the above result to hold, all that is needed is the following: (1) The more the costs of punishment decrease, the more others punish and (2) omissions are usually unintended (Dalkiran, Hoffman, Paturi, Ricketts, & Vattani, 2012; Hoffman et al., 2015).[4]

[4] In fact, even if one knows that others know that the transgression was intended, omission will still be judged as less wrong, since the transgression still won't create what game theorists call common *p*-belief, which is required for an event to influence behavior in a game with multiple equilibria.

This explanation for the omission–commission distinction leads to two novel predictions: First, for judgments and emotions not evolved to motivate witnesses to punishment but to, say, motivate witnesses to avoid dangerous partners (such as the emotion of fear; in contrast to anger or moral disgust), the omission–commission distinction is expected to be weaker or disappear altogether. Second, for transgressions of omission that, without any private information, can be presumed intentional (such as a mother who allows her child to go hungry or a person who does not give to a charity after being explicitly asked), we would not expect much of an omission–commission distinction in moral condemnation.

As with the all models discussed in this chapter, the game theoretic explanation for the omission–commission distinction does not rest on rational, conscious, strategic calculation. In fact, in this particular case, all reasonable evolutionary dynamic models lead away from punishing omissions. The fact that the above results do not rest on rational, strategic thinking is particularly important in this setting since there is evidence that the distinction between omissions and commissions is not determined deliberately but rather intuitively (Cushman, Young, & Hauser, 2006) and appears to be evolved (DeScioli et al., 2011) and that consciously considering what others believe is an onerous process (Camerer, 2003; Epley, Keysar, Van Boven, & Gilovich, 2004; Hedden & Zhang, 2002).

This same model can explain several other puzzling aspects of our morality. The first is the *means–by-product* distinction. This distinction has been documented in studies that ask respondents to judge the following variants of the classic "trolley" problem. In the standard trolley "switch" case (Foot, 1967), a runaway trolley is hurtling toward a group of five people. To prevent their deaths, the trolley must be switched onto a side track where it will kill an innocent bystander. In studies using this case, the vast majority of subjects choose the utilitarian option, judging it permissible to cause the death of one to save five (e.g., Cushman et al., 2006; Mikhail, 2007). In the "footbridge" variant (Thomson, 1976), the trolley is hurtling toward the group of five people, but the switch to divert it is inoperable. The only way to save the five is to push a man who is wearing a heavy backpack off a bridge onto the track, thereby slowing the trolley enough so the five can escape, but killing the man. In contrast to the standard switch version, where causing the death of one person is but a by-product of the action necessary to save five, most subjects in the footbridge case find it morally impermissible to force the man with the backpack onto the tracks (Cushman et al., 2006; DeScioli, Gilbert, & Kurzban, 2012)—that is, when the man is used as a means to saving the five—even though the consequences are the same, and the decision to act was made knowingly and deliberately in both cases.

Such effects are found in less contrived situations, as well. Consider the real-life distinction between terrorism, in which civilian casualties are used a means to a political goal, and anticipated collateral damage, which is a by-product of war, even when the same number of civilians are knowingly killed and the same political ends are desired (say increased bargaining power in a subsequent negotiation).

The explanation again uses "higher-order beliefs" and is based on the key insight in DeScioli et al. (2011) and formalized in Dalkiran et al. (2012) and Hoffman et al. (2015): When the harm is done as a by-product, the harm is not usually anticipated.

So even when a witness knows that the perpetrator anticipated the harm, the witness believes other witnesses will not be aware of this and will presume the harm was not anticipated by the perpetrators. For instance, suppose we observe Israel killing civilians as a by-product of a strategic raid on Hamas militants. Even if we knew Israel had intelligence that confirmed the presence of civilians, we might not be sure others were privy to this information. On the other hand, when the harm is done as a means, the harm must be anticipated, since otherwise the perpetrator would have no motive to commit the act. Why would Hamas fire rockets at civilian towns with no military presence if Hamas does not anticipate a chance of civilian casualties? Consequently, it is Nash equilibrium to punish harm done as a means but not harm done as a by-product.

Similar arguments can be made for why we find direct physical transgressions worse than indirect ones, a moral distinction relevant to, for instance, the United States' current drone policy. Cushman et al. (2006) found that subjects condemn pushing a man off a bridge (to stop a train heading toward five others) more harshly than flipping a switch that leads the man to fall through a trap door. Pushing the victim with a stick is viewed as intermediate in terms of moral wrongness. Such moral wrongness judgments are consistent with considerations of higher-order beliefs: When a man is physically pushed, any witness knows the pushing was intended, but when a man is pushed with a stick some might not realize this, and even those who realize it might suspect others will not. Even more so when a button is pressed that releases a trap door.

It is worth noting that the above argument does not depend on a specific model of punishment, as in DeScioli and Kurzban's (2009) Side-Taking Game. The above model also makes the two novel predictions enumerated above, but nevertheless captures the same basic insight. It is also worth noting the contrast between the above argument and that of Cushman et al. (2006) and Greene et al. (2009), whose models rest on ease of learning or ease of mentally simulating a situation. It is not obvious to us how those models would explain that the omission–commission and means–by-product distinctions seem to depend on priors or be unique to settings of coordinated punishment.

Why Morality Depends on Categorical Distinctions

We explain why our moral intuitions depends so much more strongly on whether a transgression occurred than on how much damage was caused. Our argument again uses coordinated punishment and higher-order beliefs: *When a categorical distinction is violated, you know others know it was violated*, *but this is not always true for continuous variables*.

Consider the longstanding norm against the use of chemical weapons. This norm recently made headlines when Bashar al-Assad was alleged to have used chemical weapons to kill about a thousand Syrian civilians, outraging world leaders who had

been silent over his use of conventional weapons to kill over 100,000 Syrian civilians. A Reuters/Ipsos poll at the time found that only 9 % of Americans favored intervention in Syria, but 25 % supported intervention if the Syrian government forces used chemical weapons against civilians (Wroughton, 2013). In the past, the United States has abided by the norm against the use of chemical weapons even at the expense of American lives: In WWII, Franklin D. Roosevelt chose to eschew chemical weapons in Iwo Jima even though, as his advisors argued at the time, their use would have saved thousands of American lives. It might even have been more humane than the flame-throwers that were ultimately used against the Japanese ("History of Chemical Weapons," 2013). We say that the norm against chemical weapons is a categorical norm because those who abide by it consider whether a transgression was committed (did Assad use chemical weapons?), rather than focusing entirely on how much harm was done (how many civilians did Assad kill?). Other norms are similarly categorical. For instance, in the introduction to this chapter, we noted that across cultures and throughout history, the norm against murder has always been categorical: We consider whether a life was terminated, not the loss of useful life years. Likewise, discrimination (e.g., during Jim Crow) is typically based on categorical definitions of race (the "one drop rule") and not, say, the darkness of skin tone. Human rights are also categorical. A human rights violation occurs if someone is tortured or imprisoned without trial, regardless of whether it was done once or many times and regardless of whether the violation was helpful in, say, gaining crucial information about a dangerous enemy or an upcoming terror attack. We even assign rights in a categorical way to all *Homo sapiens* and not based on intelligence, sentience, ability to feel pain, etc.

Why is it that we attend to such categorical distinctions instead of paying more attention to the underlying continuous variable? We use game theory to explain this phenomenon as follows: Suppose that two players (say, the United States and France) are playing a Coordination Game in which they decide whether to punish Syria, and each wants to sanction only if the other sanctions. We assume the United States does not want to levy sanctions unless it is confident France will as well, which corresponds to an assumption on the payoffs of the game (if we reverse this assumption, it changes one line in the proof, but not the result).

We model the underlying measure of harm as a continuous variable (in our example, it is the number of civilians killed). For simplicity, we assume this variable is uniformly distributed, which means Assad is equally likely to kill any number of people. This assumption is, again, not crucial, and we will point out the line in the proof that it affects. Importantly, we assume that players do not directly observe the continuous variable, but instead receive some imperfect signal (e.g., the United States observes the body count by its surveyors).

Imagine a norm that dictates that witnesses punish if their estimate of the harm from a transgression is above some threshold (e.g., levy sanctions against Syria if the number of civilians killed is estimated to be greater than 100,000). As it turns out, this is not a Nash equilibrium. To see why, consider what happens when the United States gets a signal right at the threshold. The United States thinks there is a

50 % chance that France's estimates are lower than its own[5] and, thus, that there is a 50 % chance that France's estimates are lower than the threshold. This further implies that the United States assesses only a 50 % chance that France levies sanctions, so the United States is not sufficiently confident that France will sanction, to make it in the United States's interest to sanction.

What we have shown so far is that for a threshold of 100,000, it is in the interest of the United States to deviate from the strategy dictated by the threshold norm when it gets a signal at the threshold. This means that 100,000 is not a viable threshold, and (since 100,000 was chosen arbitrarily) there is no Nash equilibrium in which witnesses punish if their estimate of the harm from a transgression is above some arbitrary threshold.

It should be noted that this result only requires that there are sufficiently many possibilities, not that there is in fact a continuum. Neither does it require that the distribution is uniform nor that the Coordination Game is not affected by the behavior of Assad. The only crucial assumptions are that the distribution is not too skewed and that the payoffs are not too dependent on the behavior of Assad (for details, see Dalkiran et al., 2012; Hoffman, Yoeli, & Dalkiran, 2015).

What happens if such norms are learned or evolved and subject to selection? Suppose there is a norm to attack whenever more than 100,000 civilians are killed. Players will soon realize that they should not attack unless, say, 100,100 civilians are killed. Then, players will learn not to attack when they estimate 100,200 civilians are killed and so on, indefinitely. Thus, every threshold will eventually "unravel," and no one will ever attack.[6]

Now let's consider a categorical norm, for example, the use of chemical weapons. We again model this as a random variable, though this time, the random variable can only take on two values (0 and 1), each with some probability. Again, players do not know with certainty whether the transgression occurred, but instead get a noisy signal. In our example, the signal represents France or the United States's assessment of whether Assad used chemical weapons, and there is some likelihood the assessors make mistakes: They might not detect chemical weapons when they had been used or might think they have detected chemical weapons when none had been used.

Unlike with the threshold norm, provided the likelihood of a mistaken signal is not too high, there is a Nash equilibrium where both players punish when they receive a signal that the transgression occurred. That is, the United States and France each levy sanctions if their assessors detect chemical weapons. This is because when the United States detects chemical weapons, the United States believes France

[5] This is where the assumption of a uniform distribution comes in. Had we instead assumed, for instance, that the continuous variable is normally distributed, then it would not be exactly 50–50 but would deviate slightly depending on the standard deviation and the location of the threshold. Nevertheless, the upcoming logic will still go through for most Coordination Games, i.e. any Coordination Game with risk dominance not too close to .5.

[6] As with omission, this follows from iterative elimination of strictly dominated strategies (see Hoffman et al., 2015, for details).

likely detected them and will likely levy sanctions. So the United States's best response is to levy sanctions. Similarly, if the United States does not detect chemical weapons, it expects France did not and will not levy sanctions, so the United States is better off not levying them.

This result is useful for evaluating whether it is worthwhile to uphold a norm. The Obama administration was harshly criticized for threatening to go to war after the Assad regime used chemical weapons but not earlier, although the regime had already killed tens of thousands of civilians. The model clarifies that Obama's position was not as inconsistent as his critics had charged: The norm against chemical weapons may be worth enforcing since it is sustainable, whereas norms against civilian casualties are harder to sustain and hence might not be worthwhile to enforce.

Let's return to some more of our motivating examples. Our model can explain why we define murder categorically: It is not possible to punish differently for different amount of quality life years taken, but it is possible to punish differentially for a life taken. As with omission–commission, however, we do expect sadness or grief to depend greatly on life years lost, even if the punishment or moralistic outrage will be less sensitive. This is a prediction of the model that, as far as we know, has yet to be tested.

Similarly, the "one-drop" rule is a categorical norm, so it can be socially enforced in an apartheid society. In contrast, consider a rule that advocates giving up one's seat for someone with lighter skin. Since this is based on a threshold in a continuous variable, while it might be enforceable by a unilateral authority, it cannot be enforced by "mob rule." Other forms of discrimination, such as discriminating against the less attractive, or the less tall, or the elderly, all being continuous variables, cannot be socially enforced via coordinated punishment, and hence, we expect such discrimination to be of a different form. In particular, it will not be based not on punishing violators. For example, male CEOs might still prefer young attractive female secretaries, and taller men are more likely to be hired as CEOs, not because of coordinated rewards or punishment but because those who hire the CEOs or secretaries are likely to be satisfying their own preferences or doing what they expect will lead to higher profits.

Likewise, the number of victims tortured by a regime or the number of lives saved by torturing is continuous. Thus, a regime cannot be punished by a coordinated attack by other countries or by a coordinated rebellion by its citizens based on the number of people tortured or the paucity of reasons for such torture. But, a regime can be attacked or overthrown depending on whether a physical harm was inflicted on a citizen by the state. Hence, human rights are treated as inalienable, even in the absence of an a priori justification for this nonutilitarian norm. And why are human rights ascribed to all living *Homo sapiens*? Perhaps not because of a good logical a priori argument, but simply because violations of human rights are enforceable by coordinated punishment, but no regime can be punished for harming any "person" of less than a certain degree of consciousness.

Finally, here is one last application. The model might also explain why revolutions are often caused by categorical events, such as a new tea tax or a single, widely

publicized self-immolation, and not a breach of a threshold in, say, the quality of life of citizens or the level of corruption. This explanation requires simply that we recognize revolutions as a coordination problem (as argued in Morris & Shin, 2002; Chwe, 2013), where each revolutionary chooses whether to revolt, and each is better off revolting only if sufficiently many others revolt.

Quirks of Altruism and the Repeated Prisoner's Dilemma with Incomplete Information

The Repeated Prisoner's Dilemma has famously been used as an explanation for the evolution of cooperation among non-kin (Axelrod & Hamilton, 1981; Dawkins, 2006; Pinker, 2003; Trivers, 1971). *In this section, we show how the same basic model can be used to explain many of the quirky features of our pro-social preferences and ideologies.*

Recall that in the Prisoner's Dilemma, each of two players simultaneously chooses whether to cooperate. Cooperation reduces a player's own payoffs by $c>0$ while increasing the other's payoffs by $b>c$. The only Nash equilibrium is for neither player to cooperate. In the Repeated Prisoner's Dilemma, the players play a string of Prisoner's Dilemmas. That is, after the players play a Prisoner's Dilemma, they learn what their opponent did and play another Prisoner's Dilemma against the same opponent with probability δ (and the game ends with probability $1-\delta$). As is well known in the evolutionary literature, there are equilibria in which players end up cooperating, provided $\delta>c/b$. In all such equilibria, cooperation is sustained because any defection by one player causes the other player to defect. This is called reciprocity. As the reader is surely familiar, there is ample evidence for the Repeated Prisoner's Dilemma as a basis for cooperation from computer simulations (e.g., Axelrod, 1984) and animal behavior (e.g., Wilkinson, 1984). The model can be extended to explain contributions to public goods if, after deciding whether to contribute to a public good, players play a Repeated Prisoner's Dilemma (see, e.g., Panchanathan & Boyd, 2004) (Fig. 5).

The key to understanding these quirks is that players often have incomplete information. For example:

1. Players do not always observe contributions. It is intuitive that, for cooperation to occur in equilibrium, contributions need to be observed with sufficiently high probability.
2. Others cannot always tell whether a player had an opportunity to contribute. For defection to be penalized, it must be the case that others can tell that a player had the opportunity to cooperate and did not (i.e. the player should not be able to hide the fact that there was an opportunity to cooperate).
3. Sometimes, there are two ways to cooperate, and one has a higher benefit, b. Then, the only way this more effective type of cooperation can be sustained in equilibrium is if others know which cooperative act is more effective.

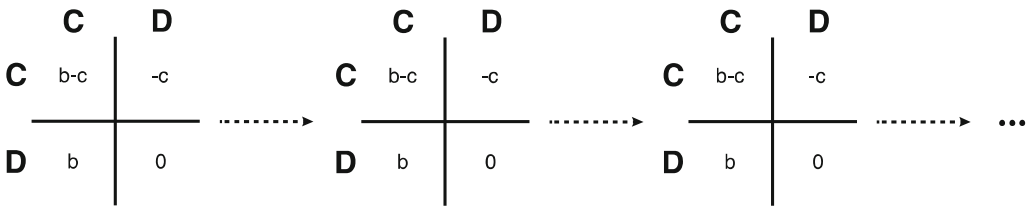

Fig. 5 The Repeated Prisoner's Dilemma. Two players play a Prisoner's Dilemma. They each observe the other's action, then, with probability δ, play another Prisoner's Dilemma against the same opponent, etc.

Technically, for the second and third point, what is needed is common knowledge that a player had an opportunity to cooperate or of the more effective means of cooperation. If observers were to know one purposely chose to defect or chose the less cooperative act, but they do not know that others know this, then observers think others will think punishment is not warranted, and observers will not punish. The argument is analogous to the discussion of higher-order beliefs in the omission–commission subsection and formalized in Dalkiran et al. (2012) and Hoffman et al. (2015).

Interpreting the Quirks of Altruism

Below we discuss some of the quirky features of altruism identified by economists and psychologists. In each case, we will argue that these features might be puzzling, but not when viewed through the lens of the above model:

Insensitivity to Effectiveness. We are surprisingly insensitive to the impact of our charitable contributions. We vote because we "want to be a part of the democratic system," or we "want to make a difference," despite the fact that our likelihood of swinging an election (even in a swing state) is smaller than our likelihood of being struck by lightning (Gelman, Silver, & Edlin, 2012). Why is our desire to "make a difference" or "be a part of the system" immune to the actual difference we are making? Our charitable contributions or volunteer efforts suffer from the same insensitivity. Why does anyone give money or volunteer time to Habitat for Humanity? The agency flies high earners who have never held a hammer halfway across the world to build houses that would be substantially more cheaply built by local experts funded by the high earners. Experimental evidence demonstrates our insensitivity: Experimental subjects are willing to pay the same amount to save 2000, 20,000, or 200,000 birds (Desvousges et al., 2010). Likewise, when donors are told their donations will be matched, tripled, or quadrupled, they donated identical amounts (Karlan & List, 2006). Why do we give so much, but do not ensure our gifts have a large impact?

The explanation follows directly from the above model: It is often the case that observers do not know which acts are effective and which are not and, certainly, this

usually is not commonly known. Thus, they will not reward or punish based on effectiveness, and we ourselves will not attend to effectiveness in equilibrium. This explanation suggests that if we want to increase efficacy of giving, we ought to focus on making sure donors' friends and colleagues are aware of the efficacy of different options. In fact, this is perhaps more important than informing the donor of efficacy, since the donor will be motivated to uncover efficacy herself.

Magnitude of the Problem. We are surprisingly unaware of and unaffected by the magnitudes of the problems we contribute to solving. How many of those who participated in the recent ALS Ice Bucket Challenge have even the vaguest sense of the number of ALS victims? (Answer: about 1/100th the victims of heart disease.) How much happier would these individuals have been if the number of ALS victims were cut in half? Multiplied by 100? The same questions could be asked about AIDS or cleft lips. If we were actually motivated by our desire to rid the world of such afflictions as we often proclaim, then we would be happier if there were fewer afflicted individuals and less happy if there were more. But we are not even aware of these numbers, let alone affected by them. This suggests an alternative motivation than the one we proclaim.

On the other hand, if we give in order to gain social rewards, it does not matter whether the problem is large or small, provided others recognize it as a problem and the social norm is to give. If our learned or evolved preferences were drastically impacted by the magnitude of the crises, we would be sensitive to whether the problem was solved, perhaps motivating us to ensure that others solve it, which we would not get credit for, or perhaps motivating us to devote too much of our resources to solving it, beyond what we would actually get rewarded for.

Observability. There is overwhelming evidence that people give more when their gifts are observed. Much of this evidence comes from the lab, where it has been demonstrated a myriad of ways (e.g., Andreoni & Petrie, 2004; Bolton, Katok, & Ockenfels, 2005; List, Berrens, Bohara, & Kerkvliet, 2004). For instance, when participants play a public goods game in the laboratory for money, their contributions are higher when they are warned that one subject will have to announce to the room of other participants how much they contributed (List et al., 2004). However, evidence also comes from real-world settings, which find large effects in settings as diverse as blood donation (Lacetera & Macis, 2010), blackout prevention (Yoeli, Hoffman, Rand, & Nowak, 2013), and support for national parks (Alpizar, Carlsson, & Johansson-Stenman, 2008). In Switzerland, voting rates fell in small communities when voters were given the option to vote by mail (Funk, 2006), which makes it harder to tell who did not vote, even though it also makes it easier to vote. In fact, our willingness to give more when observed extends to subtle, subconscious cues of being observed: People give twice as much in dictator games when there are markings on the computer screen that vaguely represent eyes (Haley & Fessler, 2005), and they are more likely to pay for bagels in their office when the payment box has a picture of eyes above it (Bateson, Nettle, & Roberts, 2006).

These results should not surprise anyone who believes our pro-social tendencies are influenced by reputational concerns (though the magnitudes are surprisingly large).

The effectiveness of subconscious cues of observability points to a primary role for reputations in our learned or evolved proclivities toward pro-social behavior. The large impact of subtle cues of observability, however, calls into question alternative explanations not based on reputations.

Explicit Requests. When we are asked directly for donations, we give more than if we are not asked, even though no new information is conveyed by the request. In a study of supermarket shoppers around Christmas time, researchers found that passersby were more likely to give to the Salvation Army if volunteers not only rang their bell but explicitly asked for a donation (Andreoni, Rao, & Trachtman, 2011). If our motive is to actually do good, or perhaps proximally to feel good by the act of giving, we should not be impacted by an explicit request.

However, if we evolved or learned to give in order to gain rewards or avoid punishment as described above, then we ought to be more likely to give when, if we did not give, it would be common knowledge that we had the option to give and chose not to. The explicit request makes the denial common knowledge.

It is worth emphasizing that our evolved intuition to respond to explicit asks may be (mis)applied to individual settings that lack social rewards. Imagine you are approached by a Salvation Army volunteer in front of a store in a city where you are visiting for one day only. A literal reading of the model would suggest that you should be no more likely to respond to an explicit request. But it is more realistic to expect that if your pro-social preferences were learned or evolved in repeated interactions then applied to this new setting, you would respond in a way that is not optimal for this particular setting and nonetheless give more when explicitly asked (just as our preferences for sweet and fatty foods, which evolved in an environment where food was scarce, lead us to overeat now that food is abundant).

Avoiding Situations in Which We Are Expected to Give. In the same supermarket study, researchers discovered that shoppers were going out of their way to exit the store through a side door, to avoid being asked for a contribution by the Salvation Army volunteers. In another field experiment, those who were warned in advance that a solicitor would come to the door asking for charitable donations were more likely to not be home. The researchers estimated that among those who gave, 50 % would have avoided being home if warned in advance of the solicitor's time of arrival (DellaVigna et al., 2012). In a laboratory analog, subjects who would have otherwise given money in a $10 dictator game were willing to pay a dollar to keep the remaining nine dollars and prevent the recipient from knowing that a dictator game could have been played (Dana et al. 2006). If our motive were to have an impact, we would not pay to avoid putting ourselves in a situation where we could have such an impact. Likewise, if our motive were to feel good by giving, we would not pay to avoid this feeling.

In contrast, if we evolved or learned to give in order to gain rewards or avoid punishment, then we would pay to avoid situations where we are expected to give. Again, this would be true even if, in this particular setting, we were unlikely to actually be punished.

Norms. People are typically *conditionally cooperative*, meaning that they are willing to cooperate more when they believe others contribute more. For example, students asked to donate to a university charity gave 2.3 percentage points more when told that others had given at a rate of 64 % than when they were told giving rates were 46 % (Frey & Meier, 2004). Hotel patrons were 26 % more likely to reuse their towels when informed most others had done the same (Goldstein, Cialdini, & Griskevicius, 2008). Households have been shown to meaningfully reduce electricity consumption when told neighbors are consuming less, both in the United States (Ayres, Raseman, & Shih, 2012) and in India (Sudarshan, 2014).

Such conditional cooperation is easily explained by the game theory model: When others give, one can infer that one is expected to give and may be socially sanctioned if one does not.

Strategic Ignorance. Those at high risk of contracting a sexually transmitted disease (STD) often go untested, presumably because if they knew they had the STD, they would feel morally obliged to refrain from otherwise desirable activity that risks spreading the STD. Why is it more reproachable to knowingly put a sexual partner at risk when one knows one has the STD than to knowingly put a sexual partner at risk by not getting tested? There is evidence that we sometimes pursue *strategic ignorance* and avoid information about the negative consequences of our decisions to others. When subjects are shown two options, one that is better for themselves but worse for their partners and one that is worse for themselves but better for their partners, many choose the option that is better for their partners. But, when subjects must first press a button (at no cost) to reveal which option is better for their partners, they choose to remain ignorant and simply select the option that is best for themselves (Dana, Weber, & Kuang, 2007).

This quirk of our moral system is again easy to explain with the above model. Typically, information about how one's actions affect others is hard to obtain, so people cannot be blamed for not having such information. When one can get such information easily, others may not know that it is easy to obtain and will not punish anyone who does not have the information. For example, although it is trivially easy to look up charities' financial ratings on websites like charitynavigator.org, few people know this and *could* negatively judge those that donate without first checking such websites. And even when others know that one can get this information easily, they might suspect that others do not know this, and so avoid punishing, since others won't expect punishment. To summarize, strategic ignorance prevents common knowledge of a violation and so is likely to go unpunished. We again emphasize that we will be lenient of strategic ignorance, even when punishment is not literally an option.

Norm of Reciprocity. We feel compelled to reciprocate favors, even if we know that the favors were done merely to elicit reciprocation and even if the favor asked in return is larger than the initial one granted (Cialdini 2001). For instance, members of Hare Krishna successfully collect donations by handing out flowers to disembarking passengers at airports, even though passengers want nothing to do with the flowers: They walk just a few feet before discarding them in the nearest bin.

Psychologists and economists sometimes take this "norm" as given, without asking where it comes from, and a naive reading of Trivers would lead one to think that we should be sensitive to the magnitude of the initial favor and whether it is manipulative.

However, according to the above model, reciprocity is the Nash equilibrium, even if the favors are not evenly matched or manipulative, since, in equilibrium, we are neither sensitive to such quantitative distinctions nor to whether the initial reciprocity was manipulative, unless these facts are commonly known.

Self-Image Concerns. People sometimes play mental tricks in order to appear *to themselves* as pro-social. For example, in an experiment, subjects will voluntarily take on a boring task to save another subject from doing it, but if given the option of privately flipping a coin to determine who gets the task, they often flip—and flip, and flip again—until the "coin" assigns the task to the other subject (Batson, Kobrynowicz, Dinnerstein, Kampf, & Wilson, 1997). Why would we be able to fool ourselves and not, say, recognize that we are gaming the coin flip? Why do we care what we think of ourselves at all? Are there any constraints on how we will deceive ourselves?

Such self-image considerations can be explained by noting that our self-image can act as a simple proxy, albeit an imperfect one, for what others think of us, and also that we are more convincing to others when we believe something ourselves (Kurzban, 2012; Trivers, 2011). This explanation suggests that the ways we deceive ourselves correspond to quirks described throughout this section—for example, we will absolve ourselves of remaining strategically ignorant even when it is easy not to, or be convinced that we have done good by voting, even if we cannot swing an election.

Framing Effects. Whether we contribute is highly dependent on the details of the experiment, such as the choice set (List, 2007) and the labels for the different choices (Ross & Ward, 1996; Roth, 1995). Such findings are often taken as evidence that social preferences cannot be properly measured in the lab (Levitt & List, 2007).

We believe a more fruitful interpretation is simply that the frame influences whether the laboratory experiment "turns on" our pro-social preferences, perhaps by simulating a situation where cooperation is expected (Levitt & List, 2007).

One-Shot Anonymous Giving: We give in anonymous, one-shot settings, such as dictator games. We also sacrifice for others in the real world when there is no chance of reciprocation: Heroes jump on grenades to save their fellow soldiers or block the door to a classroom with their bodies to prevent a school shooter from entering (Rand & Epstein, 2014). This is often seen as evidence for a role of group selection (Fehr & Fischbacher, 2003).

However, an alternate explanation is that we do not consider the likelihood of reciprocation (Hoffman et al., 2015), as described above. To explain the laboratory evidence, there are two more possibilities. First, subjects may believe there is some chance their identity will be revealed and feel the costs of being revealed as selfish are greater than the gains from the experiment (Delton, Krasnow, Cosmides, & Tooby, 2011). Second, we again emphasize that learned or evolved preferences and ideologies are expected to be applied even in novel settings to which they are not optimized.

Conclusion

In this chapter we have showed that a single approach–game theory, with the help of evolution and learning–can explain many of our moral intuitions and ideologies. We now discuss two implications.

Group Selection. Our chapter relates to the debate on group selection, whereby group level competition and reproduction is supposed to occasionally cause individuals to evolve to sacrifice their own payoffs to benefit the group (e.g., Wilson, 2006). One of the primary pieces of evidence cited in support of group selection is the existence of human cooperation and morality (Fehr & Fischbacher, 2003; Fehr, Fischbacher, & Gächter, 2002; Gintis, Bowles, Boyd, & Fehr, 2003; Haidt, 2012; Wilson, 2010, 2012), in particular: giving in one-shot anonymous laboratory experiments, intuitively sacrificing one's life for the group (jumping on the grenade), and contributions to public goods or charity. However, we have reviewed an alternative explanation for these phenomena that does not rest on group selection. It also yields predictions about these phenomena that group selection does not, such as that people are more likely to cooperate when they are being observed and there is variance in the cost of cooperation. The approach described here also explains other phenomena, such as categorical norms and ineffective altruism. These lead to social welfare losses, which is suboptimal from the group's perspective. The categorical norm against murder, for example, leads to enormous waste when keeping alive, sometimes for years, those who have virtually no chance of a future productive life.

Admittedly, despite their inefficiencies, these moral intuitions do not rule out group selection, since group selection can be weak relative to individual selection. But it does provide a powerful argument that group selection is unnecessary for explaining many interesting aspects of human morality. It also suggests that group selection is, indeed, at most, weak. One example that makes this especially clear is discrete norms. Recall that we argued that continuous norms are not sustainable because individuals benefit by deviating around the threshold. Notice that this benefit is small, since the likelihood that signals are right around the threshold is low. Group selection could easily overwhelm the benefit one would get from deviating from this Nash equilibrium, suggesting group selection is weak (i.e. there are few group-level reproductive events, high migration rates, high rates of "mutation" in the form of experimentation among individuals, etc.).

Logical Justification of Moral Intuitions. In each of the applications above, we explained moral intuitions without referring to existing a priori logical justifications by philosophers or others. Our explanation for our sense of rights does not rely on Locke's "state of nature." No argument we gave rests on God as an orderly designer, on Platonic ideals, on Kant's concepts of autonomy and humanity, etc. What does this mean for these a priori justifications? It suggests that they are not the source of our morality and are, instead, post hoc justifications of our intuitions (Haidt, 2012).

To see what we mean, consider the following analogy. One might wonder why we find paintings and sculptures of voluptuous women beautiful. Before the

development of sexual selection theory, one might have argued that perfect spheres are some kind of Platonic solid, and inherently desirable, or that curvy hips yield golden ratios. But with our current understanding of sexual selection, we recognize that our sense of beauty has evolved and that there is no platonic sense of beauty outside of that shaped by sexual selection. Any argument about perfect spheres is unparsimonious and likely flawed. Without the help of evolution and game theory, did philosophers conjure the moral equivalents of perfect spheres and golden ratios? The state of nature, the orderly designer, Platonic ideals, autonomy, and humanity, etc.—perhaps these arguments are also unfounded and unnecessary.

References

Alpizar, F., Carlsson, F., & Johansson-Stenman, O. (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics, 92*(5), 1047–1060.

Andreoni, J., & Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics, 88*(7), 1605–1623.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal, 100*, 464–477.

Andreoni, J., Rao, J. M., & Trachtman, H. (2011). *Avoiding the ask: A field experiment on altruism, empathy, and charitable giving*. Technical report, National Bureau of Economic Research.

Axelrod, R. M. (1984). *The evolution of cooperation.* New York: Basic Books.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390–1396.

Ayres, I., Raseman, S., & Shih, A. (2012). Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. *Journal of Law, Economics, and Organization*, 2–20.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters, 2*(3), 412–414.

Batson, C. D. (2014). *The altruism question: Toward a social-psychological answer*. Hillside, NJ: Psychology Press.

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335.

Bolton, G. E., Katok, E., & Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics, 89*(8), 1457–1468.

Cain, D., Dana, J., & Newman, G. (2014). *Giving vs. giving in.* Yale University Working Paper.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Chammah, A. M., & Rapoport, A. (1965). *Prisoner's dilemma; a study in conflict and cooperation*. Ann Arbor, MI: University of Michigan.

Chwe, M. (2013). *Rational ritual: Culture, coordination, and common knowledge.* Princeton, NJ: Princeton University Press.

Cone v. West Virginia Pulp & Paper Co., 330 U.S. 212, 67 S. Ct. 752, 91 L. Ed. 849 (1947).

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308–315.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

Dalkiran, N. A., Hoffman, M., Paturi, R., Ricketts, D., & Vattani, A. (2012). Common knowledge and state-dependent equilibria. In *Algorithmic game theory* (pp. 84–95). New York: Springer.
Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes, 100*(2), 193–201.
Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*(1), 67–80.
Davies, N. B. (1978). Territorial defense in the speckled wood butterfly (pararge aegeria): The resident always wins. *Animal Behaviour, 26*, 138–147.
Dawkins, R. (2006). *The selfish gene*. Oxford, UK: Oxford University Press.
DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics, 127*(1), 1–56.
Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, 108*(32), 13335–13340.
DeScioli, P., Bruening, R., & Kurzban, R. (2011). The omission effect in moral cognition: Toward a functional explanation. *Evolution and Human Behavior, 32*(3), 204–215.
DeScioli, P., Gilbert, S. S., & Kurzban, R. (2012). Indelible victims and persistent punishers in moral cognition. *Psychological Inquiry, 23*(2), 143–149.
DeScioli, P., & Wilson, B. J. (2011). The territorial foundations of human property. *Evolution and Human Behavior, 32*(5), 297–304.
DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112*(2), 281–299.
Descioli, P., & Karpoff, R. (2014). *People's judgments about classic property law cases*. Brandeis University Working Paper.
Desvousges, W. H., Johnson, F. R., Dunford, R. W., Boyle, K. J., Hudson, S. P., Wilson, K. N., et al. (2010). *Measuring nonuse damages using contingent valuation: An experimental evaluation of accuracy*. Research Triangle Park, NC: RTI Press.
Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*(3), 327.
Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785–791.
Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*(1), 1–25.
Fisher, R. A. (1958). *The genetic theory of natural selection*. Mineola, NY: Dover.
Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions.* New York: WW Norton & Co.
Frey, B. S., & Meier, S. (2004). Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review, 94*, 1717–1722.
Funk, P. (2006). Modern voting tools, social incentives and voter turnout: Theory and evidence. In *Annual Meeting of the American Economic Association*, Chicago, IL.
Gelman, A., Silver, N., & Edlin, A. (2012). What is the probability your vote will make a difference? *Economic Inquiry, 50*(2), 321–326.
Gintis, H. (2007). The evolution of private property. *Journal of Economic Behavior & Organization, 64*(1), 1–16.
Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*(3), 153–172.
Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a view-point: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research, 35*(3), 472–482.
Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.
Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York: Vintage.
Haley, K. J., & Fessler, D. M. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior, 26*.

Hedden, T., & Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games. *Cognition, 85*(1), 1–36.
Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences, 112*(6), 1727–1732.
Hoffman, M., Yoeli, E., & Dalkiran, A. (2015). *Social applications of common knowledge*. Harvard University Working Paper.
Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy, 98*, 1325–1348.
Kant, I. (1997/1787). Groundwork of the metaphysics of morals (1785). *Practical Philosophy, 108*.
Karlan, D., & List, J. A. (2006). *Does price matter in charitable giving? Evidence from a large-scale natural field experiment*. Technical report, National Bureau of Economic Research.
Kurzban, R. (2012). *Why everyone (else) is a hypocrite: Evolution and the modular mind.* Princeton, NJ: Princeton University Press.
Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization, 76*(2), 225–237.
Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives, 21*, 153–174.
List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*(3), 482–493.
List, J. A., Berrens, R. P., Bohara, A. K., & Kerkvliet, J. (2004). Examining the role of social isolation on stated preferences. *American Economic Review, 94*, 741–752.
Locke, J. (1988/1688). *Locke: Two treatises of government student edition*. Cambridge, UK: Cambridge University Press.
Mankiw, N. G. (2007). *Principles of economics*. Thomson Learning.
Milgrom, P., & Weber, R. J. (1982). The value of information in a sealed-bid auction. *Journal of Mathematical Economics, 10*(1), 105–114.
Morris, S., & Shin, H. S. (2002). Social value of public information. *The American Economic Review, 92*(5), 1521–1534.
Myerson, R. B. (2004). Justice, institutions, and multiple equilibria. *Chicago Journal of International Law, 5*, 91.
Neuwirth, R. (2005). *Shadow cities: A billion squatters, a new urban world*. New York: Routledge.
Nowak, M. A. (2006). *Evolutionary dynamics: Exploring the equations of life*. Cambridge, MA: Harvard University Press.
Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature, 432*(7016), 499–502.
Pinker, S. (2003). *The blank slate: The modern denial of human nature*. UK: Penguin.
Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences, 105*(3), 833–838.
Rand, D. G., & Epstein, Z. G. (2014). *Risking your life without a second thought: Intuitive decision-making and extreme altruism*. Available at SSRN 2424036.
Ross, L., & Ward, A. (1996). Naive realism: Implications for social conflict and misunderstanding. In T. Brown, E. Reed, & E. Turiel (Eds.), *Values and knowledge* (pp. 103–135).
Roth, A. E. (1995). Bargaining experiments. In J. H. Kagel & E. R. Alvin (Eds.), *The handbook of experimental economics* (pp. 253–342). Princeton, NJ: Princeton University Press.
Sigg, H., & Falett, J. (1985). Experiments on respect of possession and property in hamadryas baboons (*Papio hamadryas*). *Animal Behaviour, 33*(3), 978–984.
Smith, J. M., & Price, G. (1973). The logic of animal conflict. *Nature, 246*, 15.
Sosis, R., & Alcorta, C. (2003). Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology: Issues, News, and Reviews, 12*(6), 264–274.
Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*(1), 76–105.

Sudarshan, A. (2014). *Nudges in the marketplace: Using peer comparisons and incentives to reduce household electricity consumption*. Technical report, Harvard University Working Paper.
Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization, 1*(1), 39–60.
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217.
Tirole, J. (1988). *The theory of industrial organization*. Cambridge, MA: MIT Press.
Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life.* New York: Basic Books.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*, 35–57.
Trivers, R. L. (1974). Parent-offspring conflict. *American Zoologist, 14*(1), 249–264.
Wilkinson, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature, 308*(5955), 181–184.
Wilson, D. S. (2010). *Darwin's cathedral: Evolution, religion, and the nature of society*. Chicago: University of Chicago Press.
Wilson, E. O. (2012). *The social conquest of earth*. New York: WW Norton & Company.
Wilson, D. S. (2006). Human groups as adaptive units: Toward a permanent consensus. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Culture and cognition*. Oxford, UK: Oxford University Press.
Winter, E. (2014). *Feeling smart: Why our emotions are more rational than we think.* Public Books.
Wroughton, L. (2013). As Syria War Escalates, Americans Cool to U.S. Intervention: Reuter/Ipsos Poll. *Reuters*, August 24, 2013. Accessed at http://www.reuters.com/article/2013/08/25/us-syria-crisis-usa-poll-idUSBRE97O00E20130825 on 12 March 2015.
Yoeli, E., Hoffman, M., Rand, D. G., & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences, 110*(Suppl. 2), 10424–10429.
Zahavi, A. (1975). Mate selection: A selection for a handicap. *Journal of Theoretical Biology, 53*(1), 205–214.