Adaptive foundations of heroism:

Social heuristics push advantageous everyday ethical behavior to heroic extremes

Gordon T. Kraft-Todd & David G. Rand

Department of Psychology, Yale University

Appears in *Handbook of Heroism and Heroic Leadership* (2016) Eds. Allison ST, Goethals G, Kramer R. London, UK: Routledge Publishing. ISBN 978113891565. https://www.taylorfrancis.com/books/e/9781317426110

We argue that heroism is typically adaptive everyday ethical behavior taken to the extreme by over-generalization. We discuss three types of ethical principles with the properties of being cooperative, adaptive in the context of everyday life, but not in one's self-interest when taken to the extreme: justice—behaviors concerned with fairly distributing resources; solidarity—behaviors concerned with group-beneficial self-sacrifice; and pacifism—behaviors concerned with the avoidance of harming others. Because these behaviors are typically individually adaptive, they become automatized as social heuristics. These heuristics may then be (mis)applied in settings where cooperating is long-run costly to the individual, resulting in heroism.

The potential to be a hero is in all of us. This is no mere platitude; it reflects a deep and simple truth: heroism is an extreme form of everyday ethical behavior. Heroes, by our definition, are people who make great personal sacrifices for the benefit of others. From an evolutionary perspective, heroes have a mysterious origin story: how did the "self-interested" process of natural selection give rise to self-sacrifice? Extreme self-sacrificing behavior seems particularly maladaptive: wouldn't heroes have died out by now? We will argue that this puzzle can be resolved by considering the evolutionary logic supporting milder forms of self-sacrifice, the sort of ethics we see in our everyday lives.

We will specifically discuss three types of ethical principles: justice, solidarity, and pacifism. By *justice,* we mean behaviors concerned with fairly distributing resources; by *solidarity,* we mean behaviors concerned with group-beneficial self-sacrifice; and by *pacifism,* we mean behaviors concerned with the avoidance of harming others. These three ethics are not meant to be an exhaustive list, but rather a sample of key ethical principles with the properties of being: a) cooperative (i.e. individually costly but beneficial to others); b) adaptive (i.e. individually long-run payoff maximizing) in the context of everyday life; but c) not in one's self-interest (and therefore "heroic") when taken to the extreme.

There is something seemingly counterintuitive about the evolutionary nature of our thesis: how could costly other-benefiting behavior have evolved through a process that is inherently self-interested? The answer has to do with the timescale of evolution. Self-interested behaviors are often thought about in the short term: if I steal money from you today, I will be richer. It is important, however, to consider both the short- *and* long-term consequences of behavior: if I steal from you today, I will be richer today, but you will avoid me in the future and I will miss out on joint ventures with you that will make me richer in the long-term. Not stealing from you, therefore, may *seem* virtuous, but really it is in my long-term self-interest not to do so. If I started as a thief and learned the hard way, it may be that I stop stealing from you not because I am concerned about *you*, but because I am concerned about *myself*. It is therefore important to consider the consequences of behaviors not only in the short-term but in the long-term as well.

Evolution is generally considered in biological terms, as a process that involves environmental and social pressures selecting for adaptive traits which arise from randomly varying genes. But for humans, evolution can occur in the domain of culture, wherein natural selection acts not on genes, but on "memes", or units of culture such as rituals, behaviors, symbols, and strategies (Dawkins, 2006; Richerson & Boyd, 2008). Cultural evolution works via "social learning": people imitate the actions and beliefs of those whom are seen as successful. Therefore, memes that cause their adopters to succeed (i.e. that increase cultural "fitness") will spread through the population. We will discuss the evolution of heroic ethics mainly in terms of learning, though we will provide evidence of their biological origins as well.

We adopt the dual process model of cognition (Kahneman, 2003; Sloman, 1996), and argue that *social heuristics*, or rules of thumb for thinking about social interaction, play a key role in the transition from adaptive everyday ethicality to heroism. Heuristics can increase long-run payoff-maximizing behavior because they avoid the time and cognitive cost of deliberating when in familiar situations (Gigerenzer, Todd, & Group, 1999; Gilovich, Griffin, & Kahneman, 2002). Applying the general idea of heuristics to social interaction leads to the concept of cooperative intuitions that we develop through our everyday social interactions–because cooperation typically benefits our long-term self-interest (Bear & Rand, 2016; Rand et al., 2014). While cooperative intuitions may typically be adaptive—because the (often long-term) benefits generally outweigh the short-term costs—there are situations in which this is not the case. Also, because social heuristics are automatic, non-reflective processes, they are less sensitive to contextual complexity; that is, they respond similarly in situations where cooperation is advantageous as well as those where it is not. When cooperative intuitions are applied in non-advantageous contexts (i.e. where the short-term costs outweigh the long-term benefits), we often see the resulting behavior as heroic. In other words, everyday ethics are self-sacrificial helping behaviors that benefit the actor in the long run, whereas heroism is self-sacrificial helping behavior that *does not* benefit the actor in the long run.

Heroism, we therefore argue, occurs when cooperative intuitions are over-generalized to situations where they are net costly (for the individual hero). One might object, however, that not all heroism is intuitive in nature—some heroism is quite deliberate. Indeed, we identify two

types of heroism: *emergent* and *sustained*. Emergent heroism is an act of self-sacrifice that occurs "without a second thought," exemplified by many of the people honored as Carnegie Heroes who jump in front of trains or into raging rivers to save strangers (Rand & Epstein, 2014). This type of heroism arises directly from the intuitive over-generalization logic laid out above.

Sustained heroism, on the other hand, involves a long-term, often life-long commitment to self-sacrifice, exemplified by the "moral saints" of history such as Mother Teresa or Mahatma Gandhi. We believe this type of heroism could also arise from the logic above, but via a less direct route: it may be that heroic, individually costly cooperative *goals* are set via automatic processes, and this goal setting is based on over-generalization, but then deliberative processes are recruited to pursue these goals (Cushman & Morris, 2015). In other words, it could be that automatic processes lead to the establishment of (heroic) extreme goals, and then deliberative processes enact these goals, leading to sustained heroism over time. For example, many sustained heroes have a distinct moment of inspiration, as Mother Teresa did on a retreat in 1946 (Langford, 2008). It could be that her concept of justice changed to a (heroic) extreme in this moment, driven by an intuitive process, and that the rest of her life was effectively spent pursuing this heroic goal. Alternatively, it could be that deliberative processes contribute to sustained heroism by anticipating guilt for not acting in accordance with intuitive responses favoring extreme prosociality. In other words, it could be that when we deliberate and evaluate our cooperative intuitions, we consider not only the material costs and benefits of cooperating, but also the future psychological costs imposed on us by our intuitive cognitive processes: for example, we might anticipate that we would feel guilty if we do not behave cooperatively (Battigalli & Dufwenberg, 2007). Thus, whether heroism is emergent or sustained, i.e. enacted via intuition or deliberation, it may be that over-generalized cooperative intuitions are at its root.

Importantly, the claim we make here is descriptive, rather than normative. We do not argue that heroism is *bad* because it is individually non-advantageous; on the contrary, we hold that the source of heroism's virtue—the reason it is so *good*—is that it is so remarkably selfless (i.e. costly to the individual). Indeed, exploring the evolutionary roots of moral psychology (Kurzban, 2015) to discover when seemingly altruistic behavior pays off in the long-run (e.g. by benefitting one's reputation, as in Yoeli, Hoffman, Rand, and Nowak (2013)) allows us to identify when cooperation actually is extraordinary and worthy of being called *heroic*.

In sum: we argue that the everyday ethics of justice, solidarity, and pacifism arise from adaptive mechanisms. When these cooperative behaviors are typically individually adaptive, they become automatized as social heuristics. The automaticity of social heuristics makes cooperative behavior prone to over-generalization, and when this occurs in individually costly contexts, the resulting behavior is heroic. Having laid this foundation for the origins of heroism, we now provide a more detailed survey of evidence that justice, solidarity, and pacifism are adaptive and automatic.

Justice

Moral codes in spiritual and secular traditions have long hailed the virtue of justice: c.800BC, from the Sanskirt epic, the Mahabharta: "As a man himself sows, so he himself reaps; no man inherits the good or evil act of another man. The fruit is of the same quality as the action" (Hopkins, 1906); to the Bible: "Evil men do not understand justice, but those who seek Yahweh understand it fully" (Proverbs 28:5, World English Bible). Social systems are built on conceptions of justice and much of the disagreement in civil society centers on what exactly justice entails. Nearly synonymous with morality itself, justice invokes the idea of right and wrong, and the appropriate respective response of reward and punishment. But justice concerns not only "just desserts" for right and wrong, but encompass more broadly the fair distribution of resources, including the allocation of rewards and punishments. We treat justice as this broad category of behaviors that is concerned with distributing resources fairly. Justice is a puzzling ethical principle to get off the ground in an evolutionary sense, though, because treating others—particularly strangers—fairly may require self-sacrifice. As we will argue, though, justice is adaptive in everyday life, it is often automatic, and in its extreme, it is heroic.

Justice is adaptive

The logic behind the adaptive value of justice is captured by the idiom "you scratch my back and I'll scratch yours," or colloquially "tit-for-tat". This logic was formalized using game theory, which demonstrates how cooperation could evolve among non-kin via a tit-for-tat strategy with initially cooperates, and then imitates its partners move in the previous interaction (Axelrod & Hamilton, 1981). The underlying principle, *direct reciprocity*, is the idea that when individuals interact repeatedly and can remember previous interactions, cooperation among non-kin can be evolutionarily stable (Trivers, 1971). In a world in which you do not see your interaction partners again (e.g. tipping a waiter while on vacation in a foreign country), it is possible to cheat (not tip) because your interaction partner has no recourse. However, in a world in which you repeatedly interact (e.g. tipping a waiter at your favorite local restaurant), it is more difficult to cheat because your interaction partner can reciprocate the behavior (e.g. spit in your food) at your next meeting. Trivers formalized the wisdom of the folk concept of tit-for-tat, demonstrating that it can account for diverse behavior among animals of all stripes, from warning cries in birds to cooperation among humans. Here, we can see how the timescale of evolutionary analysis creates a conflict: while it is our *short-term* (i.e. one-shot interaction) self-interest to defect on others in social dilemmas, it is in our *long-term* (i.e. repeated interaction) self-interest to cooperate with them. The theory of direct reciprocity has been supported by numerous experiments in humans (Dal Bo, 2005; Fudenberg, Rand, & Dreber, 2012) as well as non-human animals (e.g. food sharing among vampire bats: (G. S. Wilkinson, 1984), and grooming in primates: (Schino, 2007); though see (T. Clutton-Brock, 2009) for the limitations of direct reciprocity in explaining non-human animal cooperation).

While direct reciprocity may seem straightforwardly applicable to relationships like those found among friends or business partners (and even family), it still does not seem to extend to strangers. That is, when we know our interactions are repeated (as in friendship or business or family), we may cooperate by the logic of tit-for-tat, but what about when we interact with people with whom we are uncertain about the repeated-ness of our future interactions? How could we establish friendships and business partnerships in the first place? And why would we ever be nice to strangers with whom we know we will *not* interact again (like the foreign waiter)?

Folk wisdom again captures the answer: "your reputation precedes you." Formally, the mechanism of reputation is called "indirect reciprocity" and the logic is this: when individuals can track each other's reputations, cooperation can be in the long-term self-interest of individuals even in non-reciprocal interactions because it affects future interaction partners' behaviors (Nowak & Sigmund, 2005). In other words, even if I do not expect to interact with someone again, if someone sees me being mean to them, they might tell others about what a bad guy I am. The nonconformist might rebut: "who cares what other people think?" Well, most of the time most of us ought to, and for selfish reasons. When other people know what a bad guy I am, they might avoid interacting with me and tell others to do likewise, and thus I pay an opportunity cost in missed future joint ventures. And worse, they may seek me out to punish me! (Fehr & Fischbacher, 2004); Cooperating for this reason, selfish as it may be, could even spark positive relationships by the logic of tit-for-tat.) Here again, we see the strategic conflict of evolutionary timescale: when there is a chance that I may be observed by others, it is in my *long-term* self-interest to cooperate with non-reciprocal interaction partners, though it is still in my *short-term* self-interest to defect on them. Note that even "observation" may be indirect; that is, my reputation can be affected not only by what others' see me do, but what others say about what I did. Thus indirect reciprocity does not require another agent to observe the interaction, but merely the ability of the recipient of the interaction to communicate the actor's behavior. In a population of agents with a certain degree of memory and communicative ability (like humans), then, even private interactions can become public record, thus motivating cooperation by the power of indirect reciprocity. Indeed, there is extensive evidence of indirect reciprocity promoting cooperation in humans (e.g. (Milinski, Semmann, & Krambeck, 2002; Yoeli et al., 2013) and there has even been some evidence in non-human animals (e.g. in sparrows, (Akçay, Reed, Campbell, Templeton, & Beecher, 2010) and in capuchins, (Anderson, Takimoto, Kuroshima, & Fujita, 2013).

So far, we have seen how the mechanisms of direct and indirect reciprocity explain how it could be adaptive for people to cooperate in social dilemmas with their (reciprocal and non-) interaction partners. But consider again the case of the foreign waiter: suppose the waiter does not know anyone I may interact with in the future, does not speak my language, and does not even know who I am (I paid in cash, so there's no identifying check or credit card) – what possible reason would I have to tip him then? Upon reflection, the answer may be: "none," yet still people do (at least, it does not seem as if there is a world-wide problem of foreigners not tipping). What could explain this seemingly irrational behavior?

Justice is automatic

To the extent that justice, or distributing resources fairly, is typically adaptive, it would be efficient to not have to think about whether to act justly at every juncture. In other words, where interactions are repeated and/or reputation is known, developing a cooperative social heuristic for just behavior can save cognitive resources. Innocuous though this logic may sound, the assertion that people may be intuitively cooperative seems to fly in the face of commonly held wisdom about human nature. After all, does the Bible not speak of original sin? And evolution of "nature, red in tooth and claw" (Tennyson, 2004)? And the rational actor model of our pursuit of self-interest?

Which perspective is correct is an empirical question – and there is evidence to support the claim that justice is automatic (Zaki & Mitchell, 2013). For example, consider experiments that manipulate cognitive processing of people playing economic cooperation games. In these games, participants choose how much money to keep for themselves versus give up to be evenly distributed between themselves and others (Peysakhovich, Nowak, & Rand, 2014). To test for automatic justice, participants were made to either decide more intuitively or to deliberate more when making their decisions (e.g. by applying time pressure or enforced delay, or by having them recall a time from their life where intuitive versus deliberative thinking worked out well). Consistent with the concept of automatic justice and the existence of social heuristics, inducing participants to rely on intuition makes them more likely to cooperate and benefit others (e.g. (Rand, Greene, & Nowak, 2012; Rand et al., 2014).

In addition to social heuristics offering efficient routes to typically advantageous outcomes, there is another reason to expect justice to be automatic: reputational costs to be seen as deliberative when deciding whether to cooperate. If I see that you have to think hard before deciding to help me, it signals that next time I'm relying on you, you might come to a different decision (if the costs of helping or the benefits of betrayal turn out to be large). If you help automatically, however, then I know that you won't stop to consider the costs and benefits, and that I can trust you. This logic of "cooperating without looking" has been formalized with a game theoretic model (Hoffman, Yoeli, & Nowak, 2015), and is supported by empirical evidence that people who make moral decisions quickly are evaluated more positively than those who choose to deliberate first (Critcher, Inbar, & Pizarro, 2013). Cooperating without looking may explain some forms of emergent heroism, where people do not to deliberate over whether to engage in an extremely costly behavior on behalf of someone else. When deciding whether to jump in front of a train or bus or save someone from drowning when there is a crowd (especially of people who know each other), it may be that the reputational benefit of quick cooperation motivates heroic action.

While there is reason to believe that learning and cultural evolution play an important role in the development of social heuristics and reputational concerns, there is also some evidence that automatic justice may be hard-wired. A growing body of research demonstrates that pre-verbal infants demonstrate prosocial tendencies, including preferring nice to mean

characters (Hamlin, Wynn, & Bloom, 2007) and voluntarily providing instrumental helping behavior (Warneken & Tomasello, 2006); although when it comes to distributing actual resources, young children tend to be quite selfish (P. R. Blake & Rand, 2010; Fehr, Bernhard, & Rockenbach, 2008).

Whether a product of genetic or cultural evolution (or both), we argue that the long-held view of a selfish human nature requiring restraint to benefit others—e.g. "For the laws of nature (as justice…) of themselves, without the terror of some power, to cause them to be observed, are contrary to our natural passions, that carry us to partiality, pride, revenge and the like" (Hobbes & Curley, 1994)—might have it backwards. Instead, it appears that we may often act with an intuitive sense of justice.

Justice in the extreme

Because a justice heuristic is automatic, it is susceptible to over-generalization, or being applied in contexts where it is *not* individually advantageous – that is, situations where the costs of cooperating will not actually be recouped in the future. For example, consider Julio Diaz, who was mugged on his way home from work one day. Feeling compassion for the teenage boy mugging him, he offered up his coat and then treated the boy to dinner (*Julio Diaz*, 2008). It seems unlikely that the boy would be a good long-term interaction partner, yet Diaz's explanation perfectly follows the logic of direct reciprocity: "I don't know, I figure, you know, you treat people right, you can only hope that they treat you right." While this act demonstrates how a justice heuristic may be over-generalized in a single instance—an example of *emergent* heroism—it may also be the case that justice heuristics are over-generalized in setting goals that lead to *sustained* heroic justice. For example, consider kidney donors who, in a very extreme form of justice, fairly distribute their two healthy kidneys, keeping one for themselves and giving the other to people in danger of having none. Certainly they cannot expect reciprocation—particularly in-kind—from the beneficiaries of their sacrifice, nor can they expect other forms of (adequate) recompense, as engaging in this act itself can serious health consequences (Segev, Muzaale, Caffo, & et al., 2010). One suggestion that such heroic commitment may originate with a justice heuristic is neurological evidence that their decisions to donate are driven by automatic, rather than deliberative, processes (Marsh et al., 2014). Alternatively, consider extreme charitable givers (MacFarquhar, 2015) like Julia Wise and Jeff Kaufman who currently live on 6% of their income and give the rest to charity. They cannot reasonably expect reciprocation from the beneficiaries of their donations, and the amount they donate is so extreme that any reputational benefits are quite unlikely to outweigh the cost they are incurring. (Further, their giving principle is based on how much they need, and so they would presumably also give away any material benefit due to improved reputation.) Yet they feel compelled to give nonetheless. These are two examples of how the typically-adaptive principle of justice may be over-generalized to set heroic goals that deliberative processes are recruited to pursue. Because these acts are not in the actor's long-term self-interest, we call this ethical behavior *extreme* and the actors *heroes*.

Solidarity

The virtue of solidarity, like justice, has appeared throughout human cultures: c.500BC, Confucius: "the noble man…takes loyalty and good faith to be of primary importance, and has no friends who are not of equal (moral) caliber" (Muller, 1990); c.50BC, Cicero: "piety admonishes us to do our duty to our country or our parents or other blood relations" (Wagenvoort, 1980); the Bible: "He who pursues righteousness and loyalty finds life, righteousness and honor." (Proverbs 21:21, New American Standard Bible). Solidarity is the bedrock of loyalty to groups of all sizes and kinds, including towns, businesses, churches, and states. But how could such group loyalty have evolved by such a supposedly selfish process as natural selection? As with justice, we will argue that solidarity is adaptive in everyday life, that it has become automatic, and in its extreme, it is heroic.

Solidarity is adaptive

There are numerous direct fitness benefits that come from being a member of a group. Group living reduces risk of predation (e.g. Treherne & Foster 1980), improves the ability to defend oneself (e.g. (Bertram, 1978). Groups also create the opportunity for specialization and division of labor, allowing positively-non-zero sum gains from trade (Durkheim, 2014). Other major benefits of group living involve improved yield to foraging (e.g. (Clark & Mangel, 1986) and increasing mechanical efficiency (e.g. in movement: (Herskin & Steffensen, 1998), or in staying warm: (Andrews & Belknap, 1986). These are only a few examples of the extensive literature on the individual fitness benefits of group living in the animal literature (for a review, see (Krause & Ruxton, 2002).

Being a member of a group is clearly adaptive, but why engage in self-sacrifice on behalf of that group? We have discussed one answer already: *indirect reciprocity*; if I am seen helping someone in the group, others will think that I am a good future interaction partner (Nowak & Sigmund, 2005). And this logic doesn't only apply to dyadic interactions; it also applies to situations where an individual's action benefits the group (Milinski et al., 2002; Panchanathan & Boyd, 2004). Crucial to the indirect reciprocity explanation, however, is that individuals have information about past behavior of other agents. (Without accurate information about the reputations of others, obviously reputation systems cannot function.) To the extent that behavior is more likely to be observable (either directly or through reputation) among ingroup members, indirect reciprocity could support greater cooperation with the ingroup than the outgroup (Masuda, 2012). Furthermore, it can be adaptive to cooperate with members of your group because they are more similar to you, and therefore more likely to have the same strategy as you – leading to the evolution of ingroup bias as a form of *tag-based* cooperation (Fu et al., 2012).

We note that some have also argued that solidarity may have evolved via *group selection* (Choi & Bowles, 2007): if intergroup competition is common, then groups whose members engage in costly cooperation within the ingroup and aggression towards the outgroup can

outcompete groups that don't. However, there is a great deal of controversy regarding whether intergroup competition was intense enough over human history to actually allow selection at the level of the group to function effectively {Burnham, 2005 #14}{Williams, 1966 #3576}{West, 2007 #218}. Therefore we do not build our adaptive argument on group selection.

Instead, we conclude that solidarity is another example of sacrificing *short-term* self-interest for *long-term* self-interest. Costly behavior on behalf of ingroup members may not be reciprocated immediately, but by solidifying one's ingroup identity, one gains access to group benefits, such as avoiding predation, increasing the gains of trade and foraging, and achieving mechanical efficiency. In sum, solidarity is adaptive: while the price of group membership may be a short-term cost, it is well worth the long-term benefits of group living.

Solidarity is intuitive

Given the benefits of group living, a social heuristics perspective would predict that preference for one's group should become automatized. And indeed, the automaticity of ingroup preferences has been demonstrated repeatedly (for a review, see Hewstone, Rubin, and Willis (2002). Intergroup bias has been explored through (following distinctions made by (Mackie & Smith, 1998)) cognition/stereotyping (e.g. Hilton and Hippel (1996)), attitudes/prejudice (e.g. Allport (1979)), and behavior/discrimination (e.g. Tajfel (1982)). Because there are many reasons that people might be motivated not to report bias (e.g. social desirability: Crowne and Marlowe (1960)), a number of implicit measures of bias have been developed (e.g. the implicit associations test, Greenwald, McGhee, and Schwartz (1998)).

Although attitudes (implicit and explicit) are shaped by the accrual of experience (e.g. E. R. Smith and DeCoster (2000), there is substantial evidence that ingroup bias emerges early in development (for a review, see Dunham, Baron, and Banaji (2008). For example, newborns prefer their mothers' native language (Moon, Cooper, & Fifer, 1993), three-month-olds prefer people of the same race (Bar-Haim, Ziv, Lamy, & Hodes, 2006), and cross-culturally, six-year-olds demonstrate implicit bias favoring people of the same race (Dunham, Baron, & Banaji, 2006). Further, implicit attitudes have been shown to develop with very little experience (Otten & Moskowitz, 2000), employing "minimal group" paradigms in which individuals are randomly sorted into groups using non-meaningful distinctions (often with false feedback) such as differences in how images are perceived (Brewer, 1979).

Solidarity is intuitive: we have briefly surveyed evidence that we have implicit ingroup bias and that this emerges early in development. Everyday solidarity predicts loyalty to group members: we have automatic cognitions favoring ingroup members and adaptive reasons for cooperating with them. When this automatic solidarity is over-generalized to extremes that are not in the actor's self-interest, it is heroic.

Solidarity in the extreme

Given the multitude of benefits to group living, solidarity heuristics are adaptive in most contexts. Here, an individual paying a short-term cost on behalf of an ingroup stands to gain future benefits; thus doing so is in their long-term self-interest. One way in which solidarity may be *over*-generalized, however, is when the short-term cost is so great as to eliminate the possibility of future benefit; in other words, when the usually short-term sacrifice becomes the *ultimate* sacrifice. This type of heroism—*emergent* heroic solidarity—is often seen in military heroes who jeopardize (and sometimes knowingly sacrifice) their lives to protect their comrades, and in a larger sense, their countrypeople's way of life. In contrast to the decision to give up *resources* for the good of a nation (via, e.g. taxes or participation in the political process), which may bring future benefits (or avoid future penalties), offering one's life for the sake of one's group invites the possibility of sacrifice that can never be repaid. To promote this kind of heroic behavior, military training places a large emphasis on acting to help one's fellow soldiers automatically without deliberation, facilitating the over-generalization of solidarity to settings of ultimate sacrifice (Grossman, 2009). More acutely, consider cases of "altruistic suicide"—when a soldier jumps on a hand-grenade to save the group. Here, the decision not merely of risking death, but of almost certain death to benefit others is more common among lower-ranking servicepeople in more cohesive groups—a finding consistent with an account explaining this behavior as an overgeneralization of reputational concerns within a well-defined and tight-knit group (J. A. Blake, 1978). Heroic solidarity may also be *sustained*. For example, consider Aung San Suu Kyi, who, leading a movement to bring democratic rule to her home country of Burma, was arrested on July 20, 1989 by the military-led government. She refused freedom in exile, enduring twenty years of house arrest and numerous attempted attacks on her person in the name of her people. Two years after her release, she said, "I'm not the only one working for democracy in Burma - there are so many people who have worked for it because they believe that this is the only way we can maintain the dignity of our people." It is difficult to imagine that her people could ever repay Suu Kyi for her sacrifice, yet still she gave it willingly. Such sustained heroic solidarity may be the result of solidarity heuristics becoming over-generalized in setting the goals that deliberative processes carry out. A small sacrifice on behalf of one's group is the stuff of everyday ethics; the *ultimate* sacrifice or a lifetime of sacrifice, may have a similar evolutionary basis, but because it is *not* in the individual's long-run self-interest, it is an example of an ethical extreme, and the person who makes it, a hero.

Pacifism

Pacifism, or the aversion to causing physical harm, also has a long tradition as a virtue in human culture. For example, consider the Hippocratic oath, c.400BC, still given by American doctors today: "either help or do not harm," (Lloyd, 1983) and well as the Christian tradition, e.g.: "Love does no harm to a neighbor; therefore love is the fulfillment of the law" (Romans

13:10, New International Version). Aggressive cultures and massive wars are, of course, ubiquitous in our history too, yet still the virtue of "not harming others" remains present in diverse cultures and throughout time. In a certain way, it seems trivial to assert pacifism as a virtue, as most of us most of the time are being nonviolent. What is virtuous about pacifism, however, is when it appears in contexts in which we would expect violence, like in Gandhi's political protests where people do not defend themselves when being attacked. Of course, in most of today's large industrialized societies, there are not many situations in which we would expect violence, as violence seems to be on the decline (Pinker, 2011). But consider the early human societies, which were on a much smaller scale, and that lacked professional armies, police forces, or formal institutions to settle disputes. Or, to strip away even more potential mechanisms that could mitigate violent conflict (e.g. language, theory of mind, etc.), consider non-human animals – why might it be in their self-interest *not* to fight?

Pacifism is adaptive

When two animals fight for access to a resource, an individual does best if she goes for the resource while the other individual retreats. If both animals approach, conflict ensues, which is costly for both parties. Thus, if you think the other will approach, you should retreat; but if you think the other will retreat, you should approach (This kind of interaction is formally modelled in game theory as the "Hawk-Dove" game, a type of anti-coordination game, (J. M. Smith & Price, 1973), or "Chicken"/"Snowdrift" in the human literature, (Rapoport & Chammah, 1966). The basic model assumes the individuals have symmetrical fighting abilities and so there is no single best strategy; instead, the best strategy is *mixed*, i.e. it alternates between approaching and avoiding. In nature, however, there is variation in fighting ability, or formally: *resource holding potential* (RHP, (Parker, 1974). (Building on this, other heterogeneities have been argued to influence the outcome of conflicts, namely motivation or *resource value*, V, (Hammerstein, 1981), and *daring*, (Barlow, Rogers, & Fraley, 1986). Because there is variation in RHP (and V and daring; we'll summarize these in the following as "strength"), strong individuals could beat weak individuals in a fight, thus incurring a lower marginal cost of fighting. Still, fighting is costly, and so it would be in the interest of strong individuals to take the resources without having to fight. Thus, it can be adaptive to signal one's strength, intimidating the other party and avoiding actual conflict.

Many animal species signal their RHP rather than fight (for a review, see (Nicholas B Davies, Krebs, & West, 2012), from beetles (West-Eberhard, 1979) to narwhals (Silverman & Dunbar, 1980) to musk ox (P. F. Wilkinson & Shank, 1976). This can be accomplished by a number of means: ritualized displays, as in red deer who assess each other through a sequence of behaviors—roaring, walking in parallel and pushing antlers against each other (T. H. Clutton-Brock, Albon, Gibson, & Guinness, 1979); visual "badges" of status, as in the plumage of the Harris sparrow, where darker plumage indicates greater RHP (Rohwer & Rohwer, 1978); or through auditory cues, as in frogs and toads, whose croak frequency is determined by body size (N. B. Davies & Halliday, 1978). Humans also have reliable signals of fighting ability: anger is perceived as a credible signal of threat (Reed, DeScioli, & Pinker, 2014), and anger is more

effectively used as a signal by stronger individuals in bargaining situations (Sell, Tooby, & Cosmides, 2009). Further, people can accurately assess upper-body strength by looking at facial structure alone (Sell, Cosmides, et al., 2009).

While it is certainly in the weaker individual's interest to avoid fighting if possible (after all, they wouldn't win!), it is also in the stronger individual's interest to avoid fighting because via signaling, they can still gain the contested resources without paying the cost of fighting. Thus, pacifism, even for the strong, is adaptive. What is the short-term v. long-term tradeoff here? Weaker parties cede the contested resource, but they live to forage another day. Stronger parties do not eliminate the competition (after all, a fight would be *more* costly for the weaker party), but by doing so, also do not risk their own injury (and perhaps a David and Goliath moment).

Pacifism is automatic

If it is adaptive to *not* harm others, even among strong individuals in a competition over resources, we might expect harm aversion to be automatic. Even in the extreme case of war, many soldiers have trouble actually pulling the trigger (Grossman, 2009). Furthermore, it is enough to mentally *simulate* actions that have harmful outcomes (e.g. smacking a baby doll on a table, but *not* smacking a broom on a table) to elicit a biophysiological stress response (Cushman, Gray, Gaffey, & Mendes, 2012). And people are willing to pay more to avoid delivering electric shocks to other people than to themselves (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014).

The automaticity of pacifism may come from our capacity for empathy. There is extensive neuroimaging evidence that feeling others' pain is associated with empathy (for reviews, see (de Vignemont & Singer, 2006) and (Hein & Singer, 2008). On the flip side, deficits in empathy have been associated with violent criminal offense (Jolliffe & Farrington, 2004). Further, neurobiological studies of abnormal brain functioning confirms the association of specific brain areas with (non)empathic behavior, e.g. reduced empathy following brain injury to the right ventromedial prefrontal cortex (Shamay-Tsoory, Tomer, Berger, & Aharon-Peretz, 2003), reduced empathy in patients with Autism spectrum disorder (Baron-Cohen & Wheelwright, 2004), and in reduced empathy in psychopaths (James Richard Blair, Jones, Clark, & Smith, 1997). It has been argued that psychopaths' particular empathy deficit—decreased response to distress cues—may play a role in their disproportionately committing violent crimes (Blair, 1995).

Empathic harm aversion is not only found in the automatic processing of adults—it has also been found to emerge early in development. One-day old babies cry when they hear tape-recorded crying of other babies (e.g. (Martin & Clark, 1982). At around two months, babies begin to show emotional synchrony with their mothers during play (e.g. (Stern, 1985). At six months, babies demonstrate a preference for puppets who help (rather than harm) another puppet (Hamlin et al., 2007). Finally, preverbal children at eighteen months can infer when an adult is struggling with a physical task and offer spontaneous help (Warneken & Tomasello, 2006).

Thus we have seen that empathic harm aversion is automatic, that it is linked with affective empathy, and that it emerges early in development. When this automatic pacifism is over-generalized to extremes that are not in the actor's self-interest, it is heroic.

Pacifism in the extreme

Pacifism is adaptive in many situations, and for humans, it is automatic. But if it happens all the time without our thinking about it, how could it be heroic? Pacifism in situations of *potential* harm, when either a weak party cedes a resource to a strong party or when a strong party signals rather than fights a weaker party, do seem commonplace. Pacifism in situations of *certain* harm, however, are more remarkable. Consider for example that you are attacked and you do not respond by attacking back—here, you risk incurring a greater cost because your self-defense could precipitate the end of the fight or mitigate the ability of your opponent to harm you. This sort of pacifism, in the face of *certain* rather than *potential* harm, is heroic, especially when exercised on behalf of another.

Our automatic aversion to harming others may serve us well in most situations—whether we are weak or strong—to help us avoid harm to ourselves. But like other heuristics, a pacifism heuristic may be insensitive to context, and applied in situations where it is not in our long-term self-interest. Consider willfully incurring physical harm, especially on behalf of others; for example, when people interpose their bodies between assailants and victims, such as when Benie Kaulesar attempted to separate his friend from a man that friend was fighting, and received a fatal blow to the head. In situations of potential harm, pacifism may serve to avoid the costs of fighting, but once the fight has begun and harm is certain, pacifism may incur greater costs, and unnecessary ones to individuals outside the fight who exhibit it on behalf of others. When putting oneself in harm's way in the context of a fight, it may be that the same automatic aversion to harm that prevents us from engaging in fights in situations of potential harm prevents us from engaging in fights when harm is certain. Therefore, we might call this *emergent* heroic pacifism due to its intuitive nature. On a grander scale, there are many examples of *sustained* heroic pacifism, as in the nonviolent political resistance led by Mahatma Gandhi. Protesting British colonial rule, 2,500 Indians marched to Dharasana Salt Works on May 21, 1930 and were met by a police force of 400. "Police charged, swinging their clubs and belaboring the raiders on all sides," observed journalist Webb Miller, "The volunteers made no resistance. As the police swung hastily with their sticks, the natives simply dropped in their tracks…The watching crowds gasped, or sometimes cheered as the volunteers crumpled before the police without even raising their arms to ward off the blows." Typically, pacifism is a strategy for avoiding harm, but here, Indian protestors did not resist the use of force, at great physical risk. Here, it may be that an automatic aversion to harm has set a goal of nonviolent resistance that protestors execute recruiting more deliberative processes. Pacifism, then, is an everyday ethical principle for the weak and strong alike when harm is uncertain, but when pacifism is exhibited in the face of certain harm—especially on behalf of others—it is extreme, and the people who do it, heroic.

Conclusion

In this chapter, we have presented an adaptive theory of heroic ethics—distributing resources fairly (justice), group-beneficial self-sacrifice (solidarity), and not harming others (pacifism)—that explains heroic behavior as an extreme form of more common "everyday heroism" that all of us exhibit, and with good reason.

For each of these ethics—justice, solidarity, and pacifism—we have argued that there are adaptive reasons to engage in the associated ethical behaviors. We then surveyed evidence that cognitions supporting these behaviors are automatic, and often emerge early in development. We argue that due to the automatic nature of the cognitive processes driving these behaviors, they are less sensitive to context and thus prone to over-generalization: these behaviors sometimes get deployed in situations where it is *not* in the individual's long-run self-interest to behaving ethically. In particular, people sometimes act ethically even when it is extremely individually costly to do so, or when no future benefits exist. That is, social heuristics sometimes take ethical behavior to the extreme. It is these extreme actions that earn the title of "heroic."

One way to interpret our argument is that heroism is foolish or irrational; after all, we have argued that heroism is about applying typically advantageous behaviors to situations in which they are actually disadvantageous (in terms of the hero's personal outcome). However, that does *not* mean that the overall strategy that leads to heroism is maladaptive. It is cognitively efficient and advantageous to sometimes rely on heuristics and intuitive processes – hence their maintenance as a key piece of human cognition (as shown, for example, in evolutionary models of dual-process agents (Tomlin, Rand, Ludwig, & Cohen, 2015; Toupo, Strogatz, Cohen, & Rand, 2015)). And as discussed above, there can be reputational benefits to "cooperating without thinking". Thus it is not foolish to adopt a strategy that leads one to sometimes in engage in heroic agents – on the contrary, selection can in fact favor such strategies (Bear & Rand, 2016; Hoffman et al., 2015).

Relatedly, the large body of literature explaining other-benefiting behavior using self-long self-interest (for a review, see Rand and Nowak (2013)) has led to disillusionment over whether there is such a thing as "pure altruism" (Lichtenberg, 2010). Here, there are two important distinctions to be made. First, cooperation is (by definition) beneficial to others, and to the extent that "action benefiting others" constitutes moral goodness, cooperation is good, whether ultimately self-interested or not. Second, however, the "self-interested cooperation is not pure altruism" argument may gain some of its condemnatory strength by confusing levels of analysis. Specifically, arguments about the evolutionary mechanisms favoring cooperation are *ultimate* explanations, that is, they explain *why* cooperation could be advantageous to individuals subject to natural selection. Critics, however, may have in mind self-interested *proximate* explanations for cooperation, that is, explanations of *how* cooperation is implemented in a given situation. In other words, altruism may be tainted when the proximate mechanism is consciously self-interested, as in the politician who gives to charity knowing the reputational benefits of doing so. The self-interestedness of evolutionary ultimate mechanisms, on the other hand, is often not in conscious awareness (when cheering for one's home sports team, does the division

of labor come to mind?), and so the moral standing of cooperative behaviors motivated by them seem less in jeopardy.

We have not presented an exhaustive account of heroism, but rather a suggestion for a theory of heroism that raises interesting questions for future work. For example, there are many models that attempt to carve morality at its joints—one question that arises from our exploration is: what are other ethical principles have evolutionary origins? For example Moral Foundations Theory (Graham et al., 2011) posits five fundamental areas of moral concern—harm, fairness, ingroup, authority, and purity. While attempts have been made to link these domains of morality to their evolutionary origins (e.g. Graham et al. (2012)), an exploration of morality "from the bottom-up" that begins with adaptive challenges may refine such theories and potentially introduce new domains of moral concern. Regarding the connection of these moral domain theories to heroism, another question is: are there types of heroism not captured by these theories? And related to our argument: are there other types of heroism that result from over-generalization of evolutionarily adaptive intuitions? And are there types of heroism that do not?

In addition to these broad questions, many specific questions might be asked to follow up on our argument. For example, regarding justice, what differences are there in the psychology of distributing resources evenly with regard to time (e.g. in vampire bats, and in common accounts of direct reciprocity) as opposed to money (e.g. in the case of wealth redistribution)? Regarding solidarity, how does group membership affect perceptions of heroism? And is heroic solidarity a solely within group phenomenon or is it recognized across group boundaries? Finally, regarding pacifism, does the connection of empathy and violence also go in the positive direction; i.e. are individuals inclined toward nonviolence more empathic? And while it has been argued that nonviolent protest is more effective than violent protest because it engenders higher participation (Chenoweth & Stephan, 2011), what is the mechanism of this contagion? Zooming out again, and of great theoretical and practical interest: when does heroism inspire others toward similar behavior?

In sum, we have argued that heroism is typically adaptive everyday ethical behavior taken to the extreme by over-generalization. Short-run sacrifices on behalf of others are typically in our long-run self-interest, and so we may develop cooperative intuitions. Whether these proximately motivate our behavior or help set long-term cooperative goals, their automaticity makes them less context-dependent, occasionally resulting in cooperative behavior that is *not* in our long-run self-interest. This truly selfless behavior, then, is different from our everyday ethical behavior not in kind, but in degree. The potential to be a hero, therefore, is in all of us.

References

Akçay, Ç., Reed, V. A., Campbell, S. E., Templeton, C. N., & Beecher, M. D. (2010). Indirect reciprocity: song sparrows distrust aggressive neighbours based on eavesdropping. *Animal Behaviour, 80*(6), 1041-1047. doi:http://dx.doi.org/10.1016/j.anbehav.2010.09.009

Allport, G. W. (1979). *The nature of prejudice*: Basic books.

Anderson, J. R., Takimoto, A., Kuroshima, H., & Fujita, K. (2013). Capuchin monkeys judge third-party reciprocity. *Cognition, 127*(1), 140-146. doi:http://dx.doi.org/10.1016/j.cognition.2012.12.007

Andrews, R. V., & Belknap, R. W. (1986). Bioenergetic benefits of huddling by deer mice (Peromyscus maniculatus). *Comparative Biochemistry and Physiology Part A: Physiology, 85*(4), 775-778. doi:http://dx.doi.org/10.1016/0300-9629(86)90294-X

Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science, 211*(4489), 1390-1396. doi:10.1126/science.7466396

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and Nurture in Own-Race Face Processing. *Psychological Science, 17*(2), 159-163. doi:10.1111/j.1467-9280.2006.01679.x

Barlow, G., Rogers, W., & Fraley, N. (1986). Do Midas cichlids win through prowess or daring? It depends. *Behavioral Ecology and Sociobiology, 19*(1), 1-8. doi:10.1007/BF00303836

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *Journal of Autism and Developmental Disorders, 34*(2), 163-175. doi:10.1023/B:JADD.0000022607.19833.00

Battigalli, P., & Dufwenberg, M. (2007). Guilt in Games. *American Economic Review, 97*(2), 170-176. doi:doi: 10.1257/aer.97.2.170

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1517780113

Bertram, B. C. (1978). Living in groups: predators and prey. *Behavioural ecology*, 64-96.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: investigating the psychopath. *Cognition, 57*(1), 1-29. doi:http://dx.doi.org/10.1016/0010-0277(95)00676-P

Blake, J. A. (1978). Death by hand grenade: Altruistic suicide in combat. *Suicide and life-threatening behavior, 8*(1), 46-59.

Blake, P. R., & Rand, D. G. (2010). Currency value moderates equity preference among young children. *Evolution and Human Behavior, 31*(3), 210-218. doi:http://dx.doi.org/10.1016/j.evolhumbehav.2009.06.012

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin, 86*(2), 307.

Chenoweth, E., & Stephan, M. J. (2011). *Why civil resistance works: The strategic logic of nonviolent conflict*: Columbia University Press.

Choi, J.-K., & Bowles, S. (2007). The Coevolution of Parochial Altruism and War. *Science, 318*(5850), 636-640. doi:10.1126/science.1144237

Clark, C. W., & Mangel, M. (1986). The evolutionary advantages of group foraging. *Theoretical Population Biology, 30*(1), 45-75. doi:http://dx.doi.org/10.1016/0040-5809(86)90024-9

Clutton-Brock, T. (2009). Cooperation between non-kin in animal societies. *Nature, 462*(7269), 51-57. Retrieved from http://dx.doi.org/10.1038/nature08366

http://www.nature.com/nature/journal/v462/n7269/pdf/nature08366.pdf

Clutton-Brock, T. H., Albon, S. D., Gibson, R. M., & Guinness, F. E. (1979). The logical stag: Adaptive aspects of fighting in red deer (Cervus elaphus L.). *Animal Behaviour, 27, Part 1*, 211-225. doi:http://dx.doi.org/10.1016/0003-3472(79)90141-6

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How Quick Decisions Illuminate Moral Character. *Social Psychological and Personality Science, 4*(3), 308-315. doi:10.1177/1948550612457688

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences, 111*(48), 17320-17325. doi:10.1073/pnas.1408988111

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of consulting psychology, 24*(4), 349.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion, 12*(1), 2. Retrieved from http://psycnet.apa.org/journals/emo/12/1/2/

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences, 112*(45), 13817-13822. doi:10.1073/pnas.1506367112

Dal Bo, P. (2005). Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *The American Economic Review, 95*(5), 1591-1604. Retrieved from http://www.jstor.org/stable/4132766

Davies, N. B., & Halliday, T. R. (1978). Deep croaks and fighting assessment in toads Bufo bufo. *Nature, 274*(5672), 683-685. Retrieved from http://dx.doi.org/10.1038/274683a0

Davies, N. B., Krebs, J. R., & West, S. A. (2012). *An introduction to behavioural ecology*: John Wiley & Sons.

Dawkins, R. (2006). *The selfish gene*: Oxford university press.

de Vignemont, F., & Singer, T. (2006). The empathic brain: how, when and why? *Trends in Cognitive Sciences, 10*(10), 435-441. doi:http://dx.doi.org/10.1016/j.tics.2006.08.008

Dunham, Y., Baron, A. S., & Banaji, M. R. (2006). From American City to Japanese Village: A Cross-Cultural Investigation of Implicit Race Attitudes. *Child Development, 77*(5), 1268-1281. doi:10.1111/j.1467-8624.2006.00933.x

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences, 12*(7), 248-253. doi:http://dx.doi.org/10.1016/j.tics.2008.04.006

Durkheim, E. (2014). *The division of labor in society*: Simon and Schuster.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature, 454*(7208), 1079-1083. doi:http://www.nature.com/nature/journal/v454/n7208/suppinfo/nature07155_S1.html

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63-87. doi:http://dx.doi.org/10.1016/S1090-5138(04)00005-4

Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., & Nowak, M. A. (2012). Evolution of in-group favoritism. *Scientific reports, 2*.

Fudenberg, D., Rand, D. G., & Dreber, A. (2012). Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *American Economic Review, 102*(2), 720-749. doi:doi: 10.1257/aer.102.2.720

Gigerenzer, G., Todd, P. M., & Group, A. R. (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*: Cambridge University Press.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2012). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, Forthcoming*.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology, 101*(2), 366. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116962/pdf/nihms246870.pdf

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology, 74*(6), 1464. Retrieved from http://psycnet.apa.org/journals/psp/74/6/1464/

Grossman, D. (2009). *On killing: The psychological cost of learning to kill in war and society*: Little, Brown and Company.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557-559. doi:http://www.nature.com/nature/journal/v450/n7169/suppinfo/nature06288_S1.html

Hammerstein, P. (1981). The role of asymmetries in animal contests. *Animal Behaviour, 29*(1), 193-205. doi:http://dx.doi.org/10.1016/S0003-3472(81)80166-2

Hein, G., & Singer, T. (2008). I feel how you feel but not always: the empathic brain and its modulation. *Current Opinion in Neurobiology, 18*(2), 153-158. doi:http://dx.doi.org/10.1016/j.conb.2008.07.012

Herskin, J., & Steffensen, J. F. (1998). Energy savings in sea bass swimming in a school: measurements of tail beat frequency and oxygen consumption at different swimming speeds. *Journal of Fish Biology, 53*(2), 366-376. doi:10.1111/j.1095-8649.1998.tb00986.x

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup Bias. *Annual Review of Psychology, 53*(1), 575-604. doi:doi:10.1146/annurev.psych.53.100901.135109

Hilton, J. L., & Hippel, W. v. (1996). STEREOTYPES. *Annual Review of Psychology, 47*(1), 237-271. doi:doi:10.1146/annurev.psych.47.1.237

Hobbes, T., & Curley, E. (1994). *Leviathan: with selected variants from the Latin edition of 1668* (Vol. 8348): Hackett Publishing.

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. doi:10.1073/pnas.1417904112

Hopkins, E. W. (1906). XXI. Modifications of the Karma Doctrine. *Journal of the Royal Asiatic Society of Great Britain & Ireland (New Series), 38*(03), 581-593.

James Richard Blair, R., Jones, L., Clark, F., & Smith, M. (1997). The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology, 34*(2), 192-198. doi:10.1111/j.1469-8986.1997.tb02131.x

Jolliffe, D., & Farrington, D. P. (2004). Empathy and offending: A systematic review and meta-analysis. *Aggression and Violent Behavior, 9*(5), 441-476. doi:http://dx.doi.org/10.1016/j.avb.2003.03.001

K. Simon (Producer). (2008, March 28). *Julio Diaz* [Audio podcast]. Retrieved from https://storycorps.org/listen/julio-diaz/

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist, 58*(9), 697.

Krause, J., & Ruxton, G. D. (2002). *Living in groups*: Oxford University Press.

Kurzban, R., & DeScioli, P. (2015). Morality. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology, 2nd Edition* (Vol. 2: Integrations, pp. 770-787): Wiley.

Langford, J. (2008). *Mother Teresa's Secret Fire*: Our Sunday Visitor.

Lichtenberg, J. (2010, October 19). Is pure altruism possible? *New York Times*.

Lloyd, G., ed. (1983). *Hippocratic Writings (2nd ed.)*. London: Penguin Books.

MacFarquhar, L. (2015). *Strangers drowning : grappling with impossible idealism, drastic choices, and the overpowering urge to help*: Penguin Press.

Mackie, D. M., & Smith, E. R. (1998). Intergroup relations: insights from a theoretically integrative approach. *Psychological review, 105*(3), 499. Retrieved from http://psycnet.apa.org/journals/rev/105/3/499/

Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W., & Cardinale, E. M. (2014). Neural and cognitive characteristics of extraordinary altruists. *Proceedings of the National Academy of Sciences, 111*(42), 15036-15041. doi:10.1073/pnas.1408440111

Martin, G. B., & Clark, R. D. (1982). Distress crying in neonates: Species and peer specificity. *Developmental psychology, 18*(1), 3.

Masuda, N. (2012). Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation. *Journal of Theoretical Biology, 311*, 8-18. doi:http://dx.doi.org/10.1016/j.jtbi.2012.07.002

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the /`tragedy of the commons/'. *Nature, 415*(6870), 424-426. Retrieved from http://dx.doi.org/10.1038/415424a

http://www.nature.com/nature/journal/v415/n6870/pdf/415424a.pdf

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development, 16*(4), 495-500. doi:http://dx.doi.org/10.1016/0163-6383(93)80007-U

Muller, C. (1990). The Analects of Confucius. Retrieved from http://www.acmuller.net/con-dao/analects.html

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291-1298. Retrieved from http://dx.doi.org/10.1038/nature04131

http://www.nature.com/nature/journal/v437/n7063/pdf/nature04131.pdf

Otten, S., & Moskowitz, G. B. (2000). Evidence for Implicit Evaluative In-Group Bias: Affect-Biased Spontaneous Trait Inference in a Minimal Group Paradigm. *Journal of Experimental Social Psychology, 36*(1), 77-89. doi:http://dx.doi.org/10.1006/jesp.1999.1399

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature, 432*(7016), 499-502. doi:http://www.nature.com/nature/journal/v432/n7016/suppinfo/nature02978_S1.html

Parker, G. A. (1974). Assessment strategy and the evolution of fighting behaviour. *Journal of Theoretical Biology, 47*(1), 223-243. doi:http://dx.doi.org/10.1016/0022-5193(74)90111-8

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nat Commun, 5*. doi:10.1038/ncomms5939

Pinker, S. (2011). *The better angels of our nature: Why violence has declined* (Vol. 75): Viking New York.

Rand, D. G., & Epstein, Z. G. (2014). Risking Your Life without a Second Thought: Intuitive Decision-Making and Extreme Altruism. *PLoS ONE, 9*(10), e109687. doi:10.1371/journal.pone.0109687

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*(7416), 427-430. doi:http://www.nature.com/nature/journal/v489/n7416/abs/nature11467.html#supplementary-information

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences, 17*(8), 413-425. doi:http://dx.doi.org/10.1016/j.tics.2013.06.003

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nat Commun, 5*. doi:10.1038/ncomms4677

Rapoport, A., & Chammah, A. M. (1966). The Game of Chicken. *American Behavioral Scientist, 10*(3), 10-28. doi:10.1177/000276426601000303

Reed, L. I., DeScioli, P., & Pinker, S. A. (2014). The Commitment Function of Angry Facial Expressions. *Psychological Science, 25*(8), 1511-1517. doi:10.1177/0956797614531027

Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution*: University of Chicago Press.

Rohwer, S., & Rohwer, F. C. (1978). Status signalling in harris sparrows: Experimental deceptions achieved. *Animal Behaviour, 26*, 1012-1022. doi:http://dx.doi.org/10.1016/0003-3472(78)90090-8

Schino, G. (2007). Grooming and agonistic support: a meta-analysis of primate reciprocal altruism. *Behavioral Ecology, 18*(1), 115-120. doi:10.1093/beheco/arl045

Segev, D. L., Muzaale, A. D., Caffo, B. S., & et al. (2010). Perioperative mortality and long-term survival following live kidney donation. *JAMA, 303*(10), 959-966. doi:10.1001/jama.2010.237

Sell, A., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., & Gurven, M. (2009). Human adaptations for the visual assessment of strength and fighting ability from the body and face. *Proceedings of the Royal Society of London B: Biological Sciences, 276*(1656), 575-584. doi:10.1098/rspb.2008.1177

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences, 106*(35), 15073-15078. doi:10.1073/pnas.0904312106

Shamay-Tsoory, S., Tomer, R., Berger, B., & Aharon-Peretz, J. (2003). Characterization of Empathy Deficits following Prefrontal Brain Damage: The Role of the Right Ventromedial Prefrontal Cortex. *Cognitive Neuroscience, Journal of, 15*(3), 324-337. doi:10.1162/089892903321593063

Silverman, H. B., & Dunbar, M. J. (1980). Aggressive tusk use by the narwhal (Monodon monoceros L.). *Nature, 284*(5751), 57-58. Retrieved from http://dx.doi.org/10.1038/284057a0

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin, 119*(1), 3.

Smith, E. R., & DeCoster, J. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. *Personality and Social Psychology Review, 4*(2), 108-131. doi:10.1207/s15327957pspr0402_01

Smith, J. M., & Price, G. (1973). lhe Logic of Animal Conflict. *Nature, 246*, 15.

Stern, D. N. (1985). *The Interpersonal World of the Infant: A View from Psychoanalysis and Developmental Psychology*: Karnac Books.

Tajfel, H. (1982). SOCIAL PSYCHOLOGY OF INTERGROUP RELATIONS. *Annual Review of Psychology, 33*(1), 1. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=11268157&site=ehost-live&scope=site

Tennyson, L. A. (2004). In Memoriam AHH. *The Norton Anthology of English Literature (New York: Norton, 1962), II, 1141*.

Tomlin, D. A., Rand, D. G., Ludwig, E., & Cohen, J. D. (2015). The evolution and devolution of cognitive control: The costs of deliberation in a competitive world. *Sci. Rep., 5*, 11002.

Toupo, D. F. P., Strogatz, S. H., Cohen, J. D., & Rand, D. G. (2015). Evolutionary game dynamics of controlled and automatic decision-making. *Chaos, 25*(7), 073120. doi:doi:http://dx.doi.org/10.1063/1.4927488

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology, 46*(1), 35-57. Retrieved from http://www.jstor.org/stable/2822435

Wagenvoort, H. (1980). Cupid and Psyche. *Pietas: Selected Studies in Roman Religion, Leiden: Brill*, 84-92.

Warneken, F., & Tomasello, M. (2006). Altruistic Helping in Human Infants and Young Chimpanzees. *Science, 311*(5765), 1301-1303. doi:10.1126/science.1121448

West-Eberhard, M. J. (1979). Sexual Selection, Social Competition, and Evolution. *Proceedings of the American Philosophical Society, 123*(4), 222-234. Retrieved from http://www.jstor.org/stable/986582

Wilkinson, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature, 308*(5955), 181-184. Retrieved from http://dx.doi.org/10.1038/308181a0

Wilkinson, P. F., & Shank, C. C. (1976). Rutting-fight mortality among musk oxen on Banks Island, Northwest Territories, Canada. *Animal Behaviour, 24*(4), 756-758. doi:http://dx.doi.org/10.1016/S0003-3472(76)80004-8

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*(Supplement 2), 10424-10429. doi:10.1073/pnas.1301210110

Zaki, J., & Mitchell, J. P. (2013). Intuitive Prosociality. *Current Directions in Psychological Science, 22*(6), 466-470. doi:10.1177/0963721413492764