

Is non-consequentialism a feature or a bug?

Fiery Cushman

In Press

The Routledge Handbook of Philosophy of the Social Mind

Ed. Julian Kiverstein

1. Introduction

Human moral values mix about nine parts function with one part flop. As psychologists, most of our time is spent trying to understand the flops. It is not that we neglect the functions entirely; they are well identified in the literature. Morality can promote cooperation (Nowak, 2006), decrease violence (Pinker, 2011), foster monogamy (Henrich, Boyd, & Richerson, 2012), protect property (Maynard Smith, 1982), promote trade (Henrich et al., 2010), and even improve our personal health (Rozin, 1997). We psychologists pay homage the functions of morality with the regularity of a priest reciting an honored liturgy, and at least half the enthusiasm.

Then, of course, we flip back to the flops with the thrill of a priest hearing scandalous confessions. Famously, people say it is wrong to sacrifice a person's life to save many others (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), and they explain this judgment by saying that it is wrong to kill (Cushman, Young, & Hauser, 2006; Hauser, Cushman, Young, Jin, & Mikhail, 2007). This has all the trappings of a flop. Do these people somehow fail to notice that the bad part of killing—the dead people—is precisely the harm minimized? Or, was there some part of killing they thought to be worse than that minor detail of death? In another well-worn flop, people claim it is wrong for siblings to engage in non-procreative sex (Haidt, Koller, & Dias, 1993b). Yet, the best reason they can give is to avoid negative consequences for the child (Bjorklund, Haidt, & Murphy, 2000). Are we truly so fearful of miraculous conception? To choose a case that is less discussed case but more commonplace, when I go to Europe

I feel obligated to tip the waiters 20%. This is well above local moral norms, and I have no special fondness for European waiters. Why are my morals burning a hole in my pocket?

Cases such as these share two common themes. First, in each case a moral value is defined over some property of an action rather than the likely consequence of the action. For instance, people prioritize the injunction against killing (an action) over the minimization of death (the outcome), the injunction against incest over the stipulated harmlessness of a specific case, and the act of tipping over the expectations of a waiter. These examples illustrate an important lesson that has been carefully documented over dozens, perhaps hundreds of psychological investigations: Human moral judgment is often non-consequentialist.

But ours isn't just any old non-consequentialism, which is the second feature that these cases share in common. Ours is "near miss" non-consequentialism: Each of the cases above comes tantalizingly *close* to consequentialism, as if they were cases of intended consequentialism gone slightly awry. Killing is usually bad, tipping is usually good, and the costs of sibling incest tend to outweigh its benefits. A natural inference is that we were really designed to be consequentialist, and that deviations from this ideal are just bugs in the program. Yet, if you point out the bugs to people they most often treat them like features, doubling down on their non-consequentialist intuitions (Cushman et al., 2006; Haidt, 2001). So, which is it: Is non-consequentialism a feature or a bug? Why is near-miss non-consequentialism a pervasive feature of human psychology?

This essay reviews three kinds of explanations. First, the errors may just be errors—byproducts of an imperfect adaptive search process. Second, the errors may reflect a tradeoff between cognitive efficiency and precision. Third, the errors may reflect a strategic commitment in the context of social interaction. These possibilities are not mutually exclusive. As we shall see, however, resolving their relative explanatory power holds important implications for the very definition of morality.

1.1 What is a consequentialist moral psychology?

In moral philosophy, consequentialism holds that moral value is a strict function of states of affairs of the world. Thus, the optimal moral behavior maximizes the desirable states of affairs.

There are at least two natural ways to define consequentialism at a psychological level. One approach is to ask, "did the individual choose an action by computing the expected value of resultant states of affairs in the world?" This approach defines consequentialism in terms of the psychological process of decision-making, and so I will refer to it as *process consequentialism*.

A second approach is to ask, "did the individual engage in a value-maximizing behavior, given the information available to her at the time of action?" This approach is

not concerned with how the individual chooses behavior, but instead with the actual behavior chosen—in other words, the policy of situation-behavior mappings that the individual enacts¹. I will therefore refer to it as *policy consequentialism*.

These two senses of psychological consequentialism are independent. For instance, consider a computer program designed to play blackjack. One approach would be for the programmer to specify the optimal move given every state of play (e.g., "if a 14 is showing, hit"). If the programmer specified the optimal set of moves then the resulting program would satisfy the criteria for policy consequentialism, but it would not involve process consequentialism because nothing about the program involves computations of expected value. As we shall see, this provides an effective analogy to some evolved reflexes, culturally inherited norms, and learned stimulus-response habits.

Conversely, consider a chess player who has memorized certain heuristics concerning favorable configurations of the board (e.g., "it is good to be in control the center"). When choosing her next move the player carefully considers the sequence of subsequent moves likely to play out, and she chooses a move that maximizes the likelihood of a favorable configuration. This involves process consequentialism because the mechanism that the player relies upon involves a computation of expected value derived from consideration of likely outcomes. Yet, it does not achieve perfect policy consequentialism, defined as the policy most likely to result in winning the game. The outcomes that the player considers are heuristic approximations of the value of states of the board, rather than exactly specified calculations for each state.

If we analogize between "winning chess" and "maximizing Darwinian fitness" then cases of this latter kind are ubiquitous. For instance, our sense of taste for nutritional and calorie-rich food is an approximation of fitness enhancing outcomes, not an exactly specified calculation. Yet, it enters into psychological processes of expected value maximization (e.g., when an animal forages, or a human chooses a dish at a restaurant). As this example makes salient, policy consequentialism can only be evaluated after committing to a position on what states of affairs should be maximized (e.g., checkmate, or fitness). It also illustrates that true policy consequentialism will rarely be attained for complex tasks; a more pertinent question is how closely it is approximated.

The utility of approximating policy consequentialism has bite, however, when the task in question is "maximizing fitness." In this case policy consequentialism is simply defined as the optimally fitness-enhancing set of behaviors for each possible state of affairs. This is because natural selection itself is consequentialist, in the sense that its processes are determined exclusively by states of affairs in the world (as opposed, say, to the past causal history that brought those states of affairs about). Yet, while policy consequentialism can be defined as optimal from a fitness perspective, often process

¹ The version of policy consequentialism advocated here is not that the agent always chooses the behavior that, *ex post*, turned out to be optimal. Rather, it is that the agent always chooses the behavior that, *ex ante*, she would rationally have believed to be optimal. Thus, an agent is a policy consequentialist if she places a bet on a coin coming up heads for the 19th time out of 20 flips, even if the coin happens to come up tails.

consequentialism cannot. Process consequentialism can guarantee a decision-maker the opportunity to make the optimal decision (given its beliefs), but it makes no guarantee about how much time or effort is involved. For tasks of real-world complexity, the computational burden of exhaustively considering all possible sequences of actions usually outweighs its benefits.

We can now revisit the motivating question of this essay with greater specificity. Human moral values appear to be organized around the functional goal of maximizing certain states of affairs (e.g. the welfare of others, or one's own fitness), yet imperfectly. Why are our moral values so often a near miss to policy consequentialism, including in circumstances where a "hit" seems attainable? And, given that process consequentialism is apparently best suited to the task of achieving policy consequentialism, why do we so often rely upon mechanisms that are non-consequentialist in the moral domain?

2. Incomplete search

Morals are shaped by natural selection operating both genetically and culturally. The basic logic of natural selection is that any random variation that enhances fitness will tend to increase in prevalence—not, of course, that only the ideal forms of random variation will flourish. To return to an example from above, our taste preferences are adapted to the task of providing sufficient calories and nutrients in our diet, but they do not operate perfectly: Aspartame has a pleasing taste even though it lacks either calories or nutritive value. An ideal system of taste preferences would track caloric and nutritive value with greater fidelity, but perhaps the necessary adaptive events simply have not occurred. We are left with a system that operates pretty well but not perfectly. Similarly, human moral values may approximate fitness maximization rather than achieving its ideal form simply because the search over the landscape of possible norms has not yet uncovered the single highest peak. The adaptive search is incomplete.

This model of “incomplete search” is often invoked in the literature as part of a broader argument that our moral instincts evolved in an environment that differs from our contemporary environment. There is a direct analogy to the evolution of our tastes: The fact that our tastes are “susceptible” to aspartame is not particularly surprising because it wasn’t available in the environment in which we evolved. (More trenchantly, our prodigious appetite for *high*-calorie foods was tuned in an environment where they were hard to obtain, but may now be maladaptive in an environment where high-calorie foods are easily obtained in a short time, at a low cost, and in a convenient foil wrapper).

This logic motivates one of Greene’s (2008a) explanations for our selective sensitivity to “up close and personal” forms of physical harm:

The rationale for distinguishing between personal and impersonal forms of harm is largely evolutionary. “Up close and personal” violence has been around for a very long

time, reaching back far into our primate lineage (Wrangham and Peterson, 1996). Given that personal violence is evolutionarily ancient, predating our recently-evolved human capacities for complex abstract reasoning, it should come as no surprise if we have innate responses to personal violence that are powerful, but rather primitive. That is, we might expect humans to have negative emotional responses to certain basic forms of interpersonal violence, where these responses evolved as a means of regulating the behavior of creatures who are capable of intentionally harming one another, but whose survival depends on cooperation and individual restraint (Sober and Wilson, 1998; Trivers, 1971).

According to this element of Greene’s model, selective pressures favor the phenotype *not harming people* (at least to a first approximation—if they are innocent, etc.). Our modern capacity for complex abstract reasoning allows us to maximize realization of this phenotype with greater reliance on process consequentialism, by estimating the harm that each of our available actions would cause to a person and then selecting the harm-minimizing action. In our deep evolutionary past, however, the only way that we tended to harm other people was through “up close and personal” violence—hitting, kicking, biting etc. Although it would have been optimal to evolve an aversion to the consequences of the actions, we happen to have first evolved a different solution: To innately feel that the very actions themselves are aversive. This adaptation prevented many harmful acts in the ancestral environment. But, it is less equipped to succeed at this task in the modern world, where our actions can often cause great harm at a physical distance, after a temporal delay, or through a chain of intermediary events. Just as our evolved tastes are susceptible to burritos, our evolved morals are susceptible to bombs.

Similar arguments are invoked widely in the literature on moral psychology, both in the context of biological evolution (Slovic, 2007) and cultural evolution (Nisbett & Cohen, 1996). And, although this argument is not often applied to morals learned through direct experience (itself a kind of adaptive process), it could be. Consider the possibility that learning morals proceeds by serial hypothesis testing about the optimal behavioral policy. In this case, a person might settle on an approximate, non-consequentialist policy that works well enough before ultimately alighting on true consequentialism.

These diverse variants of the incomplete search hypothesis share two important features. First, they explain why people may sometimes define primitive moral values over *actions* (e.g., biting) rather than over *consequences* (e.g., harm). This would appear difficult to explain on the assumption that moral values are the product of natural selection, because it is only the consequences of an action that have implications for the fitness of the agent. Yet, a moral value defined over features of an action may sufficiently enhance fitness that it is favored in the absence of a more ideal variant. Second, they view non-consequentialism as fundamentally sub-optimal—as a bug, not a feature. If random variation happened to have produced the appropriate process-

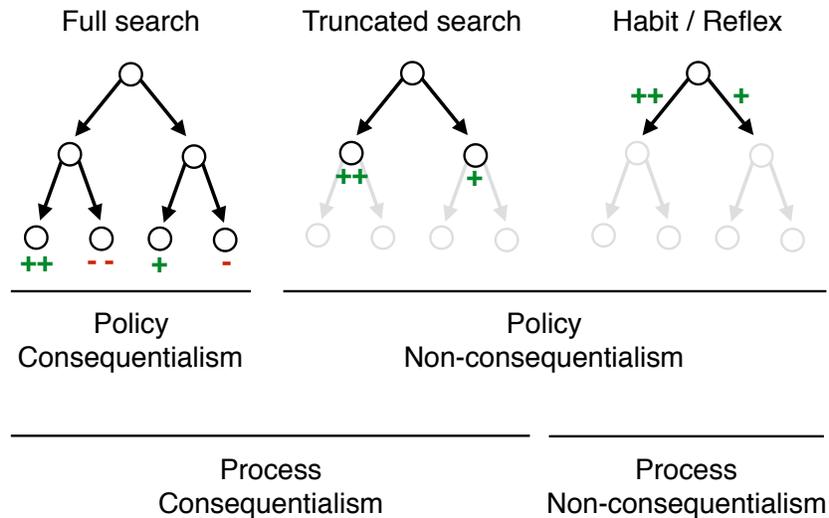
consequentialist norm, selection would have favored it. In this respect the “incomplete search” explanation of non-consequentialism differs from the other two explanations considered below.

3. Cognitive efficiency

Above we made a simplifying assumption: Because fitness depends upon the consequences of actions, an ideal cognitive mechanism for fitness maximization would select actions by considering their expected consequences. In other words, a person would choose what to do by exhaustively considering all possible outcomes of all possible actions they could perform. From the practical standpoint of information processing, however, there is nothing even remotely ideal about such a mechanism. The cognitive costs associated with this information processing task would almost certainly outweigh the benefits of subjectively optimal action selection, given the availability of a well-tuned heuristic.

The intrinsic tradeoff between computation and accuracy in decision-making has long been a foundational principle in psychology (Anderson, 1996; Thorndike, 1898) and, indeed, throughout the behavioral sciences (Kahneman, 2011). Contemporary dual process theories of cognition emphasize the competition between controlled processes, which tend to favor accuracy at the cost of cognitive effort, and automatic processes, which tend to favor cognitive efficiency at the cost of inaccuracy. Dual process models are widely applied to moral judgment and behavior (Greene, 2008b; Haidt, 2001), and it has been influentially argued that controlled processes tend to embody consequentialist principles, while automatic processes embody deontological principles. On this view, non-consequentialism is an adaptive response to information processing constraints. It is a feature, not a bug.

What, precisely, is the relationship between non-consequentialism and cognitive efficiency? Much current research seeks to identify decision-making algorithms that effectively balance computation and accuracy (Botvinick, Niv, & Barto, 2009; Daw & Shohamy, 2008; Dolan & Dayan, 2013; Sutton & Barto, 1998; Sutton, Precup, & Singh, 1999), and several of these approaches embody non-consequentialist principles. Figure 1 contrasts stylized representations of three types of approach to this problem in the form of a decision tree (of course, many other productive approaches exist). Nodes indicate decision states—discrete states of affairs for the agent. Arrows indicate actions available to the agent. Bolded areas of the decision tree are representations subject to consideration by the agent, while greyed areas of the decision tree are not explicitly considered. Thus, the greater the bolded area, the greater the cognitive demands on the agent.



“Full search” of the decision tree involves exhaustively computing the value and probabilities of all expected outcomes of every available action (Figure 1a). This approach is maximally computationally intensive, but it also guarantees optimal action selection (subject to the accuracy of the agent’s representation of the world, and the appropriateness of their assignment of reward to states of affairs).

One heuristic approach to action selection is to assign value to intermediate states that are predictive of subsequent value (Figure 1b). Above we encountered an example: A chess player might assign value to states of the board that do not correspond to checkmate, yet are associated with a relatively high probability of achieving checkmate. This allows the player to identify optimal moves without considering the full decision tree, reducing computational demands. This approach cannot guarantee optimal action selection, however, if the value assigned to intermediate states generalizes over variable circumstances. For instance, perhaps controlling the middle of the chessboard with a rook increases the probability of a subsequent checkmate for most, but not all, opponents. A player who saves cognitive effort by applying this value representation to all opponents will occasionally make suboptimal moves against the minority of opponents. Were she to instead search the full tree (with knowledge of each opponent’s strategy), she would achieve superior performance. This approach to action selection is consequentialist at a process level, but it does not achieve full policy consequentialism. Several other approaches share this characteristic. For instance, it is possible to plan towards an outcome that is correlated with but not identical to the desired outcome, as when following smoke to find fire. This corresponds to the heuristic of attribute substitution (Kahneman, 2011).

An alternative family of heuristic approaches to action selection is non-consequentialist at the process level. These work by assigning value not to states of any kind—terminal or intermediate—but to state-action pairs. In other words, given each state an agent could be in, these methods assign value directly to candidate actions that

the agent could perform. Reflexes and habitual actions have this characteristic. In each case, a basic association is encoded between a stimulus and potential behavioral responses. Although the assignment of value to state-action pairs is shaped by their historical association with consequences (for instance by natural selection, or by learning), the process of action selection proceeds without considering expected consequences. This is what makes reflexes and habits cognitively efficient.²

3.1 Innate action values

Several theories emphasize the innate assignment of moral value to specific actions or state-action pairs (Greene, 2013; Haidt, 2013; Hauser, 2006; Lieberman, Tooby, & Cosmides, 2007; Mikhail, 2011). For instance, humans have an innate aversion to physical intimacy with siblings. The adaptive value of this aversion is well understood: Genetic diversity is associated with increased fitness, and increased relatedness between parents is associated with decreased genetic diversity in their offspring. Yet, the psychological representations that give rise to our innate aversion to incest have nothing to do with these fitness costs. Rather, we assign negative value directly to the act itself.

The relevant alternative to an innate aversion to the act of incest would be innate aversion to its negative fitness consequences, together with innate knowledge of the relationship between the act and those consequences (or, on a truncated search model, some intermediary consequence such as “pregnancy resulting from incest”). This alternative mechanism is consequentialist. It would lead to qualitatively different behaviors: For instance, all else being equal, people would be comfortable engaging in non-procreative acts of intimacy with siblings. (They are not (Haidt, Koller, & Dias, 1993a; Lieberman et al., 2007; Westermarck, 1891)). Yet, it is plausible that a consequentialist adaptation to incest avoidance would not be favored over the empirically observed non-consequentialist adaptation, simply because the cost of the cognitive effort involved in computing the value of incestuous acts would not justify the negligible fitness gain (if any) of non-procreative intimacy with siblings. In other words, efficiency plausibly outweighs accuracy in this case.

Greene (2008) applies the logic of an efficiency/accuracy tradeoff to explain action-based moral principles, such as the aversion to personal harms, the attraction to spiteful revenge, and the taboo status of certain harmless actions:

Why should our adaptive moral behavior be driven by moral emotions as opposed to something else, such as moral reasoning? The answer, I believe, is that emotions are very

² In the stylized diagram above there appears to be little advantage to this approach, but in fact the advantage is large if the set of possible outcomes of any particular action is large (i.e., for stochastic state transitions). An example is the decision to hit vs. stick in blackjack. Although only two actions are available considered, hundreds of outcomes must be considered to compute the expected value of each action.

reliable and efficient responses to re-occurring situations, whereas reasoning is unreliable and inefficient in such contexts... Nature doesn't leave it to us to figure out that saving a drowning child is a good thing to do. Instead, it endows us with a powerful "moral sense" that compels us to engage in this sort of behavior (under the right circumstances). In short, when Nature needs to get a behavioral job done, it does it with intuition and emotion wherever it can. Thus, from an evolutionary point of view, it's no surprise that moral dispositions evolved, and it's no surprise that these dispositions are implemented emotionally.

The same adaptive logic that favors biological inheritance of non-consequentialist moral values may also favor the cultural inheritance of such values. For instance, it is argued that the moral norm of monogamous marriage is a cultural adaptation, and that a key function of the norm is to reduce violence by reducing the number of sexually mature single males (Henrich et al., 2012).³ Clearly, however, this consequence does not play a significant mechanistic role in the psychological processes responsible for our moral commitment to the institution of monogamous marriage. Although in this case the moral value in question is culturally acquired rather than innate, in both cases greater efficiency may be attained by stating an action-based policy rather than building in causal knowledge and relying on expected value maximization. To choose an example from outside the moral domain, cultural knowledge of CPR is typically transmitted in the form of action-based rules ("Thirty chest compressions, two breaths, repeat until medics arrive") rather than an exhaustive seminar on cardiopulmonary health. This is presumably because the efficiency benefits of simple action-based rules outweigh the theoretical and likely marginal accuracy benefits of exhaustive knowledge followed by planning. Indeed, planning in this case might actually be lethal to the patient.

A recent theoretical model formalizes this cost-benefit tradeoff in a model of the evolution of cooperation (Rand & Bear, 2016). The basic result is that, for a large parameter space, there exists a stable strategy that uniformly applies the heuristic of cooperative behavior when the costs of full deliberation are high. This leads to many instances of rational cooperation (e.g., in an iterated prisoners' dilemma) but also instances of locally irrational cooperation (e.g., in a one-shot prisoners' dilemma). When the costs of deliberation are sufficiently low, however, the individual conditions its strategy on the particular features of the game it faces.

3.2 Learned action values

Of course, there are some circumstances where it is preferable to learn a behavioral policy from experience, rather than to inherit one shaped by biological or cultural selection. This occurs, for example, when environments are not sufficiently predictable

³ This adaptive model posits selection at the level of groups rather than individuals, but this distinction does not bear on the present analysis.

on an evolutionary timescale to encode sufficiently optimal reflexes (i.e., innate stimulus-response mappings), or sufficiently accurate innate knowledge (i.e., a model from which optimal behaviors may be computed) (Littman & Ackley, 1991).

What is the relationship between learning and moral (non)-consequentialism? It is useful to begin with current psychological and neuroscientific models of value-guided learning and decision-making. These are frequently structured using the formalism of reinforcement learning, which in this case refers not to the behaviorist tradition in psychology, but rather to a branch of computer science (Sutton & Barto, 1998). Research into reinforcement learning aims to find practical algorithmic solutions to the problem of choosing an optimal policy to maximize reward in a given environment. Remarkably, this branch of computer science converged on a distinction with a longstanding and pervasive role in psychology: the distinction between planning and habit (Dickinson, Balleine, Watt, Gonzalez, & Boakes, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Thorndike, 1898). This marriage of computational and empirical approaches revolutionized the field (Dolan & Dayan, 2013; Glimcher, 2011). It also provides a natural explanation for the distinction between consequentialist and non-consequentialist moral values (Crockett, 2013; Cushman, 2013).

When an agent engages in planning, she derives the expected value of candidate actions by considering the outcomes that those actions are likely to bring about. This depends, of course, on the agent representing an internal model of the statistical associations between actions and outcomes. The agent may conduct an exhaustive search of their represented decision tree (“full planning”, illustrated in Figure 1a), or a partial search (“approximate planning”, illustrated in Figure 1b)—either way, her behavior is derived from an internally represented model of the world. In the reinforcement learning literature, therefore, such methods are referred to as *model-based*. It is easy to see how model-based reinforcement learning corresponds to process consequentialism. If social outcomes are defined as rewarding (e.g., minimizing suffering, or achieving revenge), then a model-based agent will choose actions by performing expected value computations over those outcomes.

Habits also involve an assignment of value to actions. Rather than deriving these from an internally represented model, however, these are assigned based on the history of past reward. In other words, when an agent performs an action and it leads to a favorable state of affairs, it increases the value associated with that action (specific to that context), and the opposite for actions that lead to unfavorable states of affairs. Figure 1c depicts a representation of the resulting summary of value to action. Because it does not require an internally represented model linking actions to their expected outcomes, this family of approaches is referred to as *model-free*. It has the virtue of increased computational efficiency, because the agent does not have to engage in a costly search over a statistical model of action-outcome contingencies. On the other hand, it has the drawback of inflexibility, because an agent cannot use their model to update their action values based

on model-free methods alone. Crucially, it also embodies a version of non-consequentialism. Although habit learning tends to converge to reward maximization in the long run (i.e., approximate policy consequentialism), the underlying psychological processes are non-consequentialist in the sense that they assign value directly to actions, rather than deriving those values from a model of expected outcomes.

Both of these mechanisms—model-based and model-free learning—are argued to play a key role in explaining the structure of moral cognition (Crockett, 2013; Cushman, 2013). Two elements of this proposal are provocative. The first is that it implies an overlap between the mechanisms supporting non-moral reward learning—indeed, non-*social* reward learning—and the mechanisms underlying moral judgment and behavior. In fact, much evidence supports this conclusion (Kvaran & Sanfey, 2010; Ruff & Fehr, 2014). For instance, empathy for others' pain activates similar neural substrates to the experience of pain oneself (Lamm, Decety, & Singer, 2011). Much the same is true for personally rewarding outcomes and the observation of rewards obtained by others (Braams et al., 2013; Mobbs et al., 2009). And, when people are put in the position to choose goods for others, they tend to use the same mechanisms implicated in choosing goods for themselves (Cooper, Dunne, Furey, O'Doherty, 2012; Harbaugh, Mayr, & Burghart, 2007; Shenhav & Greene, 2010), including in situations where this generosity carries a personal cost (Hare, Camerer, Knoepfle, O'Doherty, & Rangel, 2010; Janowski, Camerer, & Rangel, 2013; Zaki & Mitchell, 2011).

The second provocative element of this proposal is the claim that habit learning, in particular, can help to explain moral non-consequentialism. The concept of a “moral habit” itself has a long and distinguished philosophical pedigree, dating back at least as far as Aristotle (350/1985). Only recently, however, has the mechanism of habit learning been proposed to specifically support non-consequentialist judgments. In the trolley problem, for instance, an outcome-based assessment favors doing direct harm to a focal individual, but people find it difficult to endorse such harm. This can be understood as the consequence of negative value assigned habitually to an action: Direct, physical harm (Cushman 2013). Indeed, research has shown that people persist in their intuitive aversion to typically harmful actions (e.g., pulling the trigger of a gun pointed at a person) even when they have explicit knowledge that, in this context, the action is harmless (e.g., because it is a non-functional replica gun) (Cushman, Gray, Gaffey, & Mendes, 2012). This form of “irrational” persistence is a key signature of habit learning (Dickinson et al., 1995).

Other proposals in the literature account for moral behavior in similar manner, but without explicitly invoking the psychological concept of a habit, or the associated category of model-free reinforcement learning algorithms. For instance, the Social Heuristics Hypothesis (SHH) explains intuitive moral behavior as a consequence of reward learning mechanisms that tend to exhibit contextual inflexibility (Rand et al., 2014):

According to the SHH, people internalize strategies that are typically advantageous and successful in their daily social interactions... More reflective, deliberative processes may then override these generalized automatic responses, causing subjects to shift their behaviour towards the behaviour that is most advantageous in context. Thus, the SHH can be thought of as taking theories related to social emotions and norm internalization and making them explicitly dual process...

These approaches share the basic insight that the cognitive efficiency of non-consequentialism may outweigh its inferior accuracy.

4. Adaptive commitment

In addition to incomplete search and cognitive efficiency, a third rationale for non-consequentialist moral judgment is its potential strategic advantage in the context of social interaction. A classic illustration is the Cold War doctrine of mutual assured destruction (MAD). Nuclear states sometimes commit themselves to a policy of total and catastrophic retaliation in the event of an adversary launching a first strike. This behavior is non-consequentialist in the sense that it was widely perceived to imply the complete and mutual destruction of both participants in the conflict, whereas non-retaliation against a limited first strike could avoid complete destruction for both participants. Worse still, it could rapidly escalate a non-catastrophic strike into a catastrophic one. Yet, credible commitment to the MAD policy imparts a strategic advantage by rendering any attack by potential adversaries strictly irrational. In other words, nations achieve the best possible outcome by committing themselves to a suboptimal policy. How does this work?

The resolution to this apparent paradox lies at the intersection of a social dilemma and an inter-temporal dilemma (Rachlin, 2002). Let us make the problem concrete: President Kennedy is setting US strategic nuclear policy, and he is considering the possibility of a first strike by the USSR. In July, when Kennedy sets his policy, he is interested in setting a policy that dissuades the USSR from launching a first strike. The MAD policy is an effective deterrent if the USSR will know and trust that the US is truly committed to it. Now, suppose that Kennedy commits to this strategy in July, yet the USSR launches a first strike nevertheless in October. As the first salvo of missiles flies through the air, what is the optimal course of action for Kennedy? At this point, following through with the MAD policy is no longer preferred (assuming, as usual, that this would assure the complete destruction of both antagonists). Thus there is an inter-temporal dilemma: MAD, the policy that is optimal for Kennedy in July, is no longer optimal in October.

At first blush, it would appear that the optimal policy for the US is to *project* commitment to MAD in July, but never to act on it in the event of a future first strike.

Yet, if the USSR has the capacity for accurate inference of the US's commitment, this will not be a viable approach—the USSR will not view the US's pseudo-MAD policy as a credible threat. This is the intersecting social dilemma: The US may favor a policy that occasionally produces sub-optimal action if the expected costs are outweighed by the benefit of a social partner's trust. In other words, a rational bid to attain trust may require committing yourself to a subsequently irrational action. This occurs when social partners have the capacity to accurately assess the strength of your strategic commitments—your willingness to stand by July's logic even in a nuclear October.

How could commitment be achieved at a mechanistic level? It is clear what will not work: The organism cannot use mechanistic consequentialism (i.e., model-based reasoning) over values that align with its own fitness interests. Such a mechanism would defeat its own attempts at commitment, backing out of policies that once maximized fitness as circumstances change.

One possibility is to physically outsource commitment using a physical device, such as the “doomsday machine” envisioned in *Dr. Strangelove*. (This machine was designed to launch a retaliatory strike without the possibility of human override, foreclosing a rational override of July's decisions in October, as it were). Enforceable contracts may be used to similar effect.

Another possibility, fully implemented psychologically, is to circumvent expected value calculations in October by blindly enacting a stimulus-response routine: If a first strike is launched, reply with a retaliatory strike. This method of action selection would clearly be non-consequentialist at a process level.

A third, related approach to assuring commitment to the MAD policy would be to assign intrinsic value to “irrational” behavior: For instance, placing great positive value on the annihilation of one's antagonist and little negative value on the certain consequent annihilation of one's self. Is such a mechanism consequentialist? On the one hand, decision-making following a first strike could conform algorithmically to process consequentialism (i.e., expected value maximization) while still resulting in implementation of the MAD policy. Commitment would be achieved not by circumventing model-based reasoning, but instead by redirecting it towards the very goals of commitment. On the other hand, this approach manifestly violates policy consequentialism in that it functions precisely by *omitting* consideration of the specific outcomes most relevant to fitness—possibly including, in the case of MAD, one's own survival. This is the sense in which placing intrinsic value on commitment—for instance, the commitment to retaliation—is an adaptive form of non-consequentialism.

Employing this logic, seminal research by Frank (1988) linked the adaptive logic of non-consequentialism to emotion. Frank argued that emotions exhibit two hallmark features of adaptive non-consequentialism. First, emotions often have the consequence of compelling actions that violate local subjective consequentialism. For instance, gratitude may commit us to costly reciprocity, anger may commit us to costly acts of

spite, and the fidelity of love can commit us to the adaptive opportunity cost of foregone mating opportunities. Second, emotions tend to be observable by social partners; for instance, among people you interact with, you could probably guess who is most likely to harbor gratitude, anger or love towards you. When observable emotions credibly bind us towards non-consequentialist action, they are ideal candidates for a psychological implementation of adaptive non-consequentialism.

Current theoretical work applies this basic insight to several specific cases, providing rigorous support for the core logic based on a combination of game theory and evolutionary dynamics. For instance, in our own research we have investigated the way that it contributes to retributive punishment (Morris, McGlashan, Littman, & Cushman, in prep). We consider a setting in which one player (the thief) has the opportunity to steal from another (the victim), and then the victim has the opportunity to engage in costly punishment of the thief. Both the thief and the victim may choose their behavioral response based on a rational calculation of expected value conditioned upon their experience with their social partner, or instead based on a heritable reactive strategy. We find that adaptive processes favor flexible expected value maximization on the part of thieves, but instead favor the reactive strategy of retributive punishment on the part of victims. This can lead to functionally non-consequentialist behavior—for instance, persistent punishment of thieves who are unable to learn from such punishment. It provides an adaptive advantage, however, in that the inviolable commitment to retributive punishment drives the population equilibrium away from theft.

4.1 Reputation, partner choice and strategic non-consequentialism

The logic of adaptive non-consequentialism becomes especially powerful when combined with a family of evolutionary models involving reputation, or "partner choice". The key idea behind mutualistic partner choice is that social agents participate in an open market for cooperative partners. Put simply, because your friends have a lot of other potential friends to choose from, it makes sense to be a good friend. If you aren't good, your friends will leave you for greener pastures, and then you'll miss out on the benefits of friendship (i.e., repeated non-zero-sum exchange) (Baumard, André, & Sperber, 2013; Noë & Hammerstein, 1994; Roberts, 1998). Similar logic applies, of course, to a wide variety of social relationships. And even in mandatory social relationships, where choosing a new partner is not possible, it can be adaptive to build and maintain a positive reputation by sticking to moral commitments. This encourages investment by the mandatory social partner, to the mutual profit of both parties.

Insofar as consequentialist thought actually undermines commitment, then, the perception of consequentialist thought in others may undermine trust, reputation, friendship, and the like. Consistent with this logic, both empirical and theoretical evidence indicate that good friends act nice without thinking—i.e., for non-

consequentialist reasons. People afford greater trust, friendship and moral worth to individuals whose prosocial actions are grounded in intuitive gut feelings, compared with those whose identical prosocial actions are grounded in deliberation (Critcher, Inbar, & Pizarro, 2013; Pizarro, Uhlmann, & Salovey, 2003). Current proposals suggest that partner choice models may therefore provide an explanation for many non-consequentialist values (Baumard & Sheskin, 2015; Everett, Pizarro, & Crockett, under review). Consistent with this interpretation, people tend to make more positive personal attributions towards social partners who make characteristically deontological, versus consequentialist, moral judgments (Everett et al., under review; Uhlmann, Zhu, & Tannenbaum, 2013), and also trust them more in the context of economic exchange (Everett et al., under review).

A recent theoretical model provides a compelling explanation for these empirical phenomena in terms of partner choice (Hoffman, Yoeli, & Nowak, 2015). The model centers on a game in which a person faces the option of cooperating with or defecting against a partner in an iterated social dilemma. The allocator has a choice between considering the payoff consequences themselves on each round, or else deliberately neglecting these consequences. (Because the consequences to the self vary from round to round, cooperation will sometimes yield greater a payoff to the self, and other times yield a lesser payoff to the self). After each decision by the allocator, the responder may either continue the social interaction with the allocator or else terminate it unilaterally and irrevocably. This feature of the model corresponds to partner choice, in the sense that the social link between the players is either maintained or ceased. For a sizable parameter space, there is a subgame perfect equilibrium in which the allocator always cooperates without considering the consequences to themselves, and the responder continues the game only if the allocator maintains this behavior. Intuitively, this equilibrium is favored by the responder because it can rely on cooperation, and it is favored by the allocator because it maintains a social relationship that is, on average, profitable. The allocator's behavior is non-consequentialist at the process level because it deliberately excludes payoff-relevant information from its decision (Baumard & Sheskin, 2015).

As we have seen, the strategic equilibrium described could be realized using any of several different psychological mechanisms. One approach would be to eschew any consequence-based reasoning, and to simply assign value to the action of cooperating with social partners. This corresponds to process non-consequentialism. Another approach would be to engage in consequence-based reasoning but to systematically exclude payoffs to oneself from the reward function, maximizing exclusively on payoffs to the relevant social partners. This approach is process consequentialist. There is a mismatch, however, between the consequences that make the policy adaptively favored (positive fitness consequences for oneself) and the consequences that are maximized at the level of psychological mechanism, which exclusively apply to others. In this way, even the latter approach may be thought of as a kind of adaptive non-consequentialism.

4.2 Reputation and moral identity

As we have seen, non-consequentialism works best when it is perceived to arise from non-strategic motives. Thus you will find the best friends if you are perceived to be irrationally charitable, and not to be strategically angling for reciprocity. Likewise, you will dissuade potential adversaries best when you are perceived to be irrationally vengeful, not a strategic bluffer. On the crucial assumption is that people have some ability to perceive others' decision strategies (albeit a limited ability); this furnishes an adaptive rationale for *actual* non-consequentialist moral values, not just apparent ones.

This final point—the limitations of our ability to detect others' motives—has crucial implications for the nature of moral thought and behavior. It means that we are forever in a position of uncertainty, attempting to assess whether the social behavior we observe in others reflects a genuine and reliable commitment (i.e., process non-consequentialism), or whether instead it reflects veiled strategic motives and plans. Such worries must have weighed heavily on minds of those world leaders who controlled nuclear arsenals during the height of the cold war. They are probably just as common among anyone who has ever been to high school, fallen in love, gossiped over a water cooler, or wielded any of the other nuclear arms of our daily social lives.

This inferential problem is fraught because many moral acts can be explained equally well by two hypotheses: True commitment, or strategic feigning. When a lover gives roses on Valentine's Day, is this an expression of love or an act of manipulation? When a bully threatens, is this a manifestation of hair-trigger aggression or a shallow bluff? Often, we cannot know. There are some circumstances, however, that provide strong evidence in favor of a single hypothesis, such as when a lover binds himself to the contract of marriage (committed!), or is spied through a window hitting on his waitress (strategic!). Such episodes can trigger the reassessment of past behavior—suddenly last year's Valentine's Day roses don't look so rosy. As a consequence, when a person suspected of being strategic, it can be difficult to repair reputational damage. After all, future acts of apparent commitment may always be interpreted in light of one's past departure from the primrose path.

This introduces an imperative interest in maintaining one's moral reputation (Aquino & Reed II, 2002; Gausel & Leach, 2011; Jordan et al., 2016; Sperber & Baumard, 2012; Wojciszke, 2005), and it suggests that even small deviations from moral behavior may be catastrophic (Tannenbaum, Uhlmann, & Diermeier, 2011). It also suggests that we should have psychological mechanisms designed specifically for the task of assessing others' moral character, with a powerful purchase on our judgments and behaviors (Martin & Cushman, 2015). It may also explain why moral features are considered a particularly fundamental dimension of personal identity (Strohming & Nichols, 2014, 2015). Finally, it suggests a natural explanation for the motivating

concepts and intuitions of philosophical virtue ethics. In summary, an account of act-based non-consequentialism married with the logic of partner choice provides a natural account of person-based moral judgment (Uhlmann, Pizarro, & Diermeier, 2015).

5. Conclusion: What is morality?

I have reviewed three distinct explanations for the non-consequentialism of human moral psychology: Incomplete search, cognitive efficiency, and an adaptive commitment. According to the first model, non-consequentialism is a bug. According to the second model, it is a feature, but it carries a distinct cost. The desirable property of non-consequentialist judgments is not that they are non-consequential, but merely that they are cognitively efficient. According to the third model, however, the very property of non-consequentialism embodies a key adaptive advantage.

Presumably each of these explanations plays a role in explaining some subset of non-consequentialist morals. Future work must resolve whether one of them tends to be a more prevalent explanation than the others. Insofar as this dictates whether non-consequentialism tends more often to be a feature or a bug, it may carry implications for the normative status of non-consequentialist ethical systems.

The relative scope of each explanation has a second implication, however, that is at once less obvious and yet more certain: It may help us to define morality.

According to the models of incomplete search and cognitive efficiency, non-consequentialism is not at all special to morality. Rather, the same forces that give rise to moral non-consequentialism would be expected to give rise to non-consequentialist mechanisms in the domain of personal choice and resource maximization. In fact, in some cases the very same mechanisms might be responsible for decision-making in both domains. This encapsulates a dominant view in social and cognitive neuroscience: Moral decision-making is a subset of ordinary decision-making that involves more-or-less identical mechanisms operating in more-or-less the same neural substrates (Ruff & Fehr, 2014). At a mechanistic level, there are two features of morality that stand out as semi-distinctive. First, the core architecture of decision-making interfaces with an architecture designed to represent and reason about others' mental states (Buckner, Andrews Hanna, & Schacter, 2008; Young & Dungan, 2012). Second, the architecture operates with distinct primitive forms of reward in the moral domain—for instance, the reward of reciprocity in a state of gratitude, or of revenge in a state of fury.

At its heart, however, this perspective does not favor a definition of morality in terms of specific mechanism; rather, it invites a definition in terms of the adaptive function accomplished by relatively domain-general mechanisms. This perspective is widely adopted in the field. For instance, Haidt (2008) offers: "Moral systems are interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate

selfishness and make cooperative social life possible." He refers to this as a "functionalist" definition of morality: "Rather than specifying the content of moral issues (e.g., "justice, rights, and welfare"), this definition specifies the function of moral systems" (Haidt, 2010). Greene (2015) takes a similar approach, arguing that some natural kinds are "bound together, not at the mechanical level, but at the functional level. I believe that the same is true of morality. So far as we can tell, the field of moral cognition does not study a distinctive set of cognitive processes (Greene, 2014, but see Mikhail, 2011.) Instead, it studies a set of psychological phenomena bound together by a common function. As I (Greene, 2013) and others (Frank, 1988; Gintis, 2005; Haidt, 2012) have argued, the core function of morality is to promote and sustain cooperation."

The model of adaptive commitment invites quite a different perspective on the definition of morality, especially when combined with the logic of partner choice and the demands of maintaining a moral identity. This model furnishes a unique rationale for non-consequentialist mechanisms in moral judgment and behavior. Moreover, it predicts that people will be motivated to signal their moral identity by adhering generally to moral norms, and will be motivated to identify such signals in others. It therefore favors a definition of the moral domain that is more squarely focused on a set of psychological mechanisms than their underlying functional rationale (although certainly a coherent functional rationale exists). In short, it suggests a definition of morality centered precisely on the property of non-consequentialism.

This would be a controversial definition. For instance, it has been forcefully argued that at least some human moral judgments are basically consequentialist at a process level, and that these mechanisms are to be normatively preferred (Greene, 2013). How might this tension be resolved? One obvious solution is definitional pluralism, admitting both a functional of morality (including judgments and behaviors derived from domain-general processes) alongside a mechanistic definition (embodying the logic of non-consequentialist commitment).

How different are these definitions, in their practical effect? A long time ago, or in places far away—when social interactions were governed by kinship, reputation and coalition—the mechanisms of adaptive non-consequentialism often constituted the most effective means of reaping the adaptive benefits of cooperation. There and then, functional and mechanistic definitions of morality mostly picked out the same features of our psychology and behavior. In the contemporary West, however, social interaction is often governed by formal institutions such as states, corporations and clubs. These largely undermine the rationale for adaptive non-consequentialism. I do not have to ask about the reputation of the taxi driver whom I hail, nor does she need to ask about mine, because institutional structures align each of our interests towards a successful exchange of money for service. Here and now, the psychological mechanisms that fulfill the functional definition of morality often diverge from those that fulfill its mechanistic definition.

References

- Anderson. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, 51(4), 355.
- Aquino, & Reed II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6), 1423.
- Baumard, André, & Sperber. (2013). A mutualistic approach to morality: the evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(01), 59-78.
- Baumard, & Sheskin. (2015). 3 Partner Choice and the Evolution of a Contractualist Morality. *The Moral Brain: A Multidisciplinary Perspective*, 35.
- Bjorklund, Haidt, & Murphy. (2000). Moral dumbfounding: When intuition finds no reason. *Lund Psychological Reports*, 2, 1-23.
- Botvinick, Niv, & Barto. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3), 262-280.
- Braams, Güroğlu, de Water, Meuwese, Koolschijn, Peper, & Crone. (2013). Reward-related neural responses are dependent on the beneficiary. *Social Cognitive and Affective Neuroscience*, nst077.
- Buckner, Andrews Hanna, & Schacter. (2008). The brain's default network. *Annals of the New York Academy of Sciences*, 1124(1), 1-38.
- Cooper, Dunne, Furey, O'Doherty. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience*, 24(1), 106-118.
- Critcher, Inbar, & Pizarro. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, 4(3), 308-315.
- Crockett. (2013). Models of morality. *Trends in Cognitive Sciences*.
- Cushman. (2013). Action, Outcome, and Value: A Dual-System Framework for Morality. *Personality and Social Psychology Review*, 17(3), 273-292. doi: Doi 10.1177/1088868313495594
- Cushman, Gray, Gaffey, & Mendes. (2012). Simulating murder: The aversion to harmful action. *Emotion*, 12(1), 2-7.
- Cushman, Young, & Hauser. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, 17(12), 1082-1089.
- Daw, & Shohamy. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, 26(5), 593-620.
- Dickinson, Balleine, Watt, Gonzalez, & Boakes. (1995). Motivational control after extended instrumental training. *Learning & behavior*, 23(2), 197-206.
- Dolan, & Dayan. (2013). Goals and habits in the brain. *Neuron*, 80(2), 312-325.
- Everett, Pizarro, & Crockett. (under review). Inference of trustworthiness from intuitive moral judgments.
- Frank. (1988). *Passion Within Reason: The Strategic Role of the Emotions*. New York: Norton.
- Gausel, & Leach. (2011). Concern for self-image and social image in the management of moral failure: Rethinking shame. *European Journal of Social Psychology*, 41(4), 468-478.

- Glimcher. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3), 15647-15654.
- Greene. (2008a). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 3). Cambridge, MA: MIT Press.
- Greene. (2008b). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), (Vol. 3). Cambridge, MA: MIT Press.
- Greene. (2013). Moral Tribes: Emotion, Reason and the Gap Between Us and Them.
- Greene. (2015). The rise of moral cognition. *Cognition*, 135, 39-42.
- Greene, Sommerville, Nystrom, Darley, & Cohen. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108.
- Haidt. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814-834.
- Haidt. (2013). The righteous mind: Why good people are divided by politics and religion.
- Haidt, Koller, & Dias. (1993a). Affect, culture and morality, or, is it wrong to eat your dog? *Journal of Personalis and Social Ps\ cholog.*
- Haidt, Koller, & Dias. (1993b). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65(4), 613-628.
- Harbaugh, Mayr, & Burghart. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, 316(5831), 1622-1625.
- Hare, Camerer, Knoepfle, O'Doherty, & Rangel. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of neuroscience*, 30(2), 583-590.
- Hauser. (2006). *Moral Minds: How nature designed a universal sense right and wrong*. New York: Harper Collins.
- Hauser, Cushman, Young, Jin, & Mikhail. (2007). A dissociation between moral judgment and justification. *Mind and Language*, 22(1), 1-21.
- Henrich, Boyd, & Richerson. (2012). The puzzle of monogamous marriage. *Philosophical Transactions Of The Royal Society B-Biological Sciences*, 367(1589), 657-669.
- Henrich, Ensminger, Mcelreath, Barr, Barrett, Bolyanatz, . . . Ziker. (2010). Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science*, 327(5972), 1480. doi: 10.1126/science.1182238
- Hoffman, Yoeli, & Nowak. (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.
- Janowski, Camerer, & Rangel. (2013). Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. *Social Cognitive and Affective Neuroscience*, 8(2), 201-208.
- Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, 530(7591), 473-476.
- Kahneman. (2011). *Thinking, fast and slow*: Farrar Straus & Giroux.
- Kvaran, & Sanfey. (2010). Toward an Integrated Neuroscience of Morality: The Contribution of Neuroeconomics to Moral Cognition. *Topics in Cognitive Science*, 2(3), 579-595.

- Lamm, Decety, & Singer. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage*, 54(3), 2492-2502.
- Lieberman, Tooby, & Cosmides. (2007). The architecture of human kin detection. *Nature*, 445(7129), 727-731.
- Littman, & Ackley. (1991). *Adaptation in Constant Utility Non-Stationary Environments*. Paper presented at the ICGA.
- Martin, & Cushman. (2015). To punish or to leave: Distinct cognitive processes underlie partner control and partner choice behaviors.
- Maynard Smith. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
- Mikhail. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press.
- Mobbs, Yu, Meyer, Passamonti, Seymour, Calder, . . . Dalgleish. (2009). A key role for similarity in vicarious reward. *Science*, 324(5929), 900-900.
- Montague, Dayan, & Sejnowski. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of neuroscience*, 16(5), 1936-1947.
- Morris, McGlashan, Littman, & Cushman. (in prep). Flexible theft and resolute punishment.
- Nisbett, & Cohen. (1996). *Culture of Honor: The Psychology of Violence in the South*. Boulder: Westview Press Inc.
- Noë, & Hammerstein. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, 35(1), 1-11.
- Nowak. (2006). Five rules for the evolution of cooperation. *Science*, 314, 1560-1563.
- Pinker. (2011). *The better angels of our nature: Why violence has declined*. Penguin Books.
- Pizarro, Uhlmann, & Salovey. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, 14(3), 267-272.
- Rachlin. (2002). Altruism and self-control. *Behavioral and Brain Sciences*, xxx, yyy-zzz.
- Rand, & Bear. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*.
- Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak, & Greene. (2014). Social heuristics shape intuitive cooperation. *Nature communications*, 5.
- Roberts. (1998). Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394), 427-431.
- Rozin. (1997). Moralization. In A. Brandt & P. Rozin (Eds.), *Morality and Health*. (pp. 379-401). New York: Routledge.
- Ruff, & Fehr. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, 15(8), 549-562.
- Schultz, Dayan, & Montague. (1997). A neural substrate of prediction and reward. *Science*, 275, 1593-1599.

- Shenhav, & Greene. (2010). Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude. *Neuron*, 67(4), 667-677.
- Slovic. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision Making*, 2(2), 79-95.
- Sperber, & Baumard. (2012). Moral reputation: an evolutionary and cognitive perspective. *Mind & Language*, 27(5), 495-518.
- Strohminger, & Nichols. (2014). The essential moral self. *Cognition*, 131(1), 159-171.
- Strohminger, & Nichols. (2015). Neurodegeneration and identity. *Psychological Science*, 0956797615592381.
- Sutton, & Barto. (1998). *Introduction to reinforcement learning*: MIT Press.
- Sutton, Precup, & Singh. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1), 181-211.
- Tannenbaum, Uhlmann, & Diermeier. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, 47(6), 1249-1254.
- Thorndike. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied*, 2(4), i-109.
- Uhlmann, Pizarro, & Diermeier. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, 10(1), 72-81.
- Uhlmann, Zhu, & Tannenbaum. (2013). When it takes a bad person to do the right thing. *Cognition*, 126(2), 326-334.
- Westermarck. (1891). *The History of Human Marriage*. London: MacMillan.
- Wojciszke. (2005). Morality and competence in person-and self-perception. *European review of social psychology*, 16(1), 155-188.
- Young, & Dungan. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience*, 7(1), 1-10.
- Zaki, & Mitchell. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences*, 108(49), 19761-19766.