PUBLIC AFFAIRS QUARTERLY
Volume 24, Number 2, April 2010

PREJUDICE AND EVOLUTIONARY GAME THEORY

Malcolm Murray

THE PROBLEM

Let us define prejudice as a propensity to treat members of a particular outgroup as having less moral worth than members of one's own group. Racism and sexism are kinds of prejudice, but so, too, is homophobia, as well as some fervent nationalisms.[1] Prejudice is viewed as a problem for evolutionary ethics: prejudice clearly exists in our world, yet we also deem prejudice immoral.[2] How can an evolutionary account explain the fit of a trait x at the same time as explaining the fit of a trait that tries to expunge x? If moral traits are those with fit (as evolutionary ethicists like to imagine), why would we not say prejudice is moral, as opposed to immoral? If both prejudice and moral indignation against prejudice can be explained on evolutionary models, what use are evolutionary models? This problem is a specific application of the general descriptive-normative problem. As even one of the evolutionary game theory pioneers, Maynard Smith, observes, "A scientific theory—Darwinism or any other—has nothing to say about the value of a human being."[3] Assuming Smith is not doubting that humans have value, I believe that he is wrong here. The value of human beings is not something that we discover. Rather, we find that ascribing value to human beings confers advantage to the ascribers, and that advantage can be tracked through evolutionary models.

Evolutionary ethics explains morality in purely naturalistic terms. We can predict that moral strategies will evolve without inferring any intentionality, any objective moral values, or any supernatural forces. If morality is an evolved trait, then morality (like any adaptation) has advantage *only instrumentally*—only given certain environmental conditions, none of which apply necessarily. Some people take this to be the very problem with evolutionary models since they cannot show why morality is good *in itself*. Presumably Maynard Smith clung to this view, whereas, for me, the rejection of the antiquated notion that morality is *good in itself* is precisely what gives evolutionary ethics the edge.

In brief, my answer to the above problem is this: racism, or prejudicial at-

titudes in general, although not adaptations in their own right, piggyback on an otherwise adaptive trait. Thus, we can explain on purely evolutionary grounds why nonprejudice is moral yet why prejudice lingers. To make my case, I appeal to modern accounts of evolutionary biology, some psychological and sociological theories about prejudice, and some evolutionary games of my own making.

Success of Conditional Cooperation

In evolutionary games, strategies that conditionally correlate with one another do better than strategies that do not. Correlation takes slightly different forms in different games. In all such games, the ones prone to play well with other like-minded sorts are the ones we operationally define as moral strategies.[4] Immoral agents are the ones who like to exploit others and who fail to play well with others, including their like-minded ilk. It is precisely this reluctance to play well with their own kind that makes immoral strategies evolutionarily unstable. The more immoral sorts there are, the worse those immoral sorts do. Whereas, the more moral strategies there are, the better those moral strategies do. Despite immorality paying in the short term, morality pays over the long haul.

A propensity to correlate with others by itself is not sufficient for success, however. The correlation has to be a subset of a family of strategies we can call *conditional*. Benefits to defectors accrue only in cases where defectors are allowed to *unilaterally defect*. Only through unilateral defection can one successfully exploit a cooperator. Thus, the success of moral strategies is contingent on their recognizing and defecting against other unilateral defectors. We can see this in Axelrod's Tit-For-Tats (TFT), Gauthier's constrained maximizers, Trivers's reciprocal altruists, Danielson's reciprocal cooperators (partly), and Skyrms's fairmen, among others.[5] More specifically, however, a Conditional Cooperator (CC) is any player who cooperates with those who cooperate with her and defects against those who defect against her.[6]

A CC agent will comply with the agreements she has found reason to make when interacting with others who are also prone to comply. As well noted in the literature, these agents do better than other types of agents in terms of evolutionary fitness. They are those who can play well with others. Defectors, on the other hand, play poorly with others. The short-term benefit of being a defector means that the population of defectors grows. But as the population of defectors grows, defectors start doing worse. Whereas, the more CC agents there are, the better CC agents do, which will entail a growth in the CC population. This is why, despite the short-term rationality of defection, CCs will take over a large proportion of the population (allowing niches for rational defectors.)[7] Given the success of CC in a wide range of evolutionary games, I can narrow our focus to examine the effects of prejudice on CC agents.

Mixed Signals

Evolutionary game theory shows that conditional cooperation is a wonderfully successful and simple moral strategy. Unfortunately, conditional cooperation has an ugly side-effect. Markers of cooperative dispositions can be too easily associated with irrelevant features, like skin color, sex, sexual orientation, cultural gestures, facial features, or social space. If such differences cloud our ability to detect propensity to cooperate, the prudent resolve is to defect. Heuristic shortcuts can therefore develop, in which irrelevant differences become markers for noncorrelation. In other words, prejudice may be the progeny of moral strategies, not immoral strategies.

Game theory teaches us that one should cooperate with other cooperators and defect against other defectors. But such advice is not helpful if we cannot well discern who is a cooperator and who a defector. Whenever there is doubt, for example in initial interactions, a prudent resolve is to defect.[8] Meanwhile, detecting cooperative propensity may be influenced by cultural norms. Being unfamiliar with other cultures may preclude the ability to detect propensity for cooperation. Prejudice may be simply an overextension of invoking the default position, given mixed signals. It may well be mere opacity between races that explains our racist tendencies.[9]

The TFT and CC agents belong in the same family of conditional strategies in the sense that they both cooperate with cooperators and defect against defectors. The CC agents and TFT agents have slightly different structures, but a similar default switch can be understood with TFT agents as well.[10] A TFT strategy, *A*, will cooperate with an agent who cooperated with *A* on the previous round, and will defect with an agent who defected against *A* on the previous round. Thus, TFT strategies have the propensity of collapsing into a perpetual feud, especially when affected by prejudice. If a white, say, is less able to distinguish one black from another as she can distinguish one white from another, and has been defected against by a black, the propensity to treat any black as *that* defector increases. In James Baldwin's "Sonny's Blues," some drunken white boys decide, for fun, to drive over a black boy. The boy dies. The white kids escape in their car and are never punished, let alone sought. The boy's brother (Sonny's father) witnessed the event and "till the day he died . . . weren't sure but that every white man he saw was the man that killed his brother."[11] In this sense, the success of TFT strategies can be hampered if they, or profile-prone TFTs, adopt group-identity rather than individual identity in determining who has cooperated and who has defected in a previous round. If one is unable to decode the pre-play cues, it is not an unwise strategy to defect merely to avoid the cost of being exploited. Such behavioral trends will be indistinguishable from prejudice.

Explanations relying on misidentification also corroborate the "contact thesis" for prejudice reduction. The contact hypothesis predicts that prejudice will be

reduced with increased contact between members of the various groups.[12] For example, whites who regularly watch basketball were found to be more likely to detect individual differences in blacks than whites who do not regularly watch basketball (where basketball players are predominantly black).[13] Grim and colleagues for example, have claimed to corroborate the contact model using spatialized Prisoner Dilemma (PD) games with TFT agents.[14]

The mixed-signal hypothesis cannot capture the whole picture, however. Real persons playing PDs in experimental situations will vary their moves according to how the other *unseen* player is identified. They are more likely to offer initial cooperative moves if the other player is identified as an in-group (a teammate, for example), and more likely to offer initial defect moves if the other player is identified as a member of an out-group (someone from a rival university, for example).[15] Similarly, randomly dividing people into arbitrary teams creates non-cooperative behaviors between members of "opposing" teams. In other words, despite original contact, prejudice can gradually consume a population. This suggests that prejudicial attitudes are more hardwired, more visceral than the mixed-signal hypothesis. Mere contact does not seem to be enough. We need to feel we are also on the same team. (Admittedly, one way to make one feel that we are on the same team is through increased contact.)

Evolution and Piggybacking

If prejudice confers advantage, as seems evident given the propensity of prejudice in human history, then the evolutionary model would seem to be forced to say prejudice is moral. Similar logic led Spencer and the Social Darwinists to conclude that the moral act is to let the sickly die and the impoverished perish. It gives evolutionary ethics a bad name, to say the least. But such characterizations of evolutionary ethics are misinformed.[16] We have sentiments against culling the poor and sickly, and evolutionary ethics needs to account for such sentiments. Similarly, evolutionary ethics needs to account for anti-prejudice. I think it can if we conceive of prejudicial attitudes (including racism) as an offshoot of an otherwise adaptive trait.

Evolutionary biologists can easily accept the fact that traits that were adaptive to one environment have not survived in current environments and traits currently existing need not be adaptations. A panda has an opposable thumb on its forelimbs, and this is an adaptive trait since it confers advantage in terms of mobility and foraging. But, along with its hand thumbs, it also has foot thumbs, which are not used in any (obvious) advantageous way. Presumably, the same genetic trait that produces variation in the wrist (to produce a thumb) also produced the variation in the ankle (to produce the foot thumb).[17] In such a case, we do not say the foot thumb confers advantage. Simply, it piggybacks on the thumb that does confer advantage. The continued presence of the trait (the foot thumb) is explained by a biological constraint rather than adaptation.

If a trait is successful in one environment and can be reproduced in future generations, and the environment is relatively stable across these future generations (which includes resource availability and current predator–prey relations, etc.), then we can predict that that trait will remain fairly stable. On the other hand, precisely because environments are not stable, variability is essential to biological species, and a variation on a past successful trait may confer fitness in a new and ever-changing environment. Moreover, traits that are adaptive in very narrow conditions are not likely to confer fitness since there are low odds of those narrow conditions applying. Traits that can confer some advantage (even if less than ideal advantage) in a wider range of environmental conditions are more likely to be adaptive. But in any event, since the only traits that can evolve are predicated on existing traits, this means nature often works with what it has, rather than what would be ideal for the circumstances. This is because evolution has to work on small changes to existing traits. If an existing trait has been selected for, then changes to that trait will more likely make it maladaptive. (Tinkering with something that works is more likely to make it not work, as opposed to making it work better.)[18] A single strategy that produces wildly divergent behavioral patterns is not odd in biology.[19] Despite the vast array of diversity found in nature, nature abhors complexity in terms of its basic building blocks, as the stark anatomical similarities across species illustrates. There is no reason to think strategies concerning social interaction are different in this respect.

Evidence for heuristics abounds in nature.[20] So long as certain heuristics can develop that work in normally occurring situations, the speed and convenience of these cheap heuristics may benefit agents who have developed them more than agents who have not. The cost of finely tuned algorithms may not be sufficiently compensatory, given the environments they generally inhabit. The program in a greylag goose to retrieve stray eggs to her nest can get fooled by footballs, skulls, and lightbulbs.[21] The drive for male red-winged blackbirds to defend their territory is triggered not merely by the red wings of invading male red-winged blackbirds, but also by red balls, red shirts, red balloons, etc.

This is (or may be) the same with morality. Given the success of CC, what we define as the moral strategy is the propensity to cooperate with cooperators and defect against defectors. But variation can exist concerning what clues we use to interpret the presence of a defector. The more lax we are, the more likely that others will exploit us. The more stringent we are, the fewer chances we will have to reap mutual benefit. Ignoring psychological and sociological explanations, we can account for the variability concerning prejudicial attitudes in purely probabilistic terms.

Racism, then, and prejudice in general, may well be a carryover of an otherwise adaptive trait. Like the panda's foot thumb, prejudice may persist—not because it confers fitness, but because it is correlated with moral behavior that confers fitness.

Game Theory Method

Players and Scoring. Following on the success of CC in other evolutionary games, I assume all players are CC agents. That is, in a prisoner's dilemma, they cooperate with those prone to cooperate with them and defect against those prone to defect against them. I have divided these CC agents first into two groups differentiated by a color marking: Whites (W) and Blacks (B). I have further divided these two groups into nonprejudice (N) and prejudice (P). The Ns cooperate with any cooperator, whether W or B. The Ps cooperate with others of their same color only. Thus, a WP cooperates with WN and WP, but not BN or BP. Given this basic structure, there is a question concerning whether a prejudiced individual gets to exploit a nonprejudiced individual or not. Let us distinguish, at least initially, between an exploitation game and a non-interaction game. In the exploitation game, a WP defects against a BN, while the BN cooperates with WP. Thus, WP reaps an advantage at BN's expense. Such a possibility may seem odd when we are dealing with CC agents only, but if racists are being pitted against pure CC agents, they won't survive. To account for the lingering of racists amid the growth of CC agents, we have to conceive of racists as being a subset of CC—ones who cooperate with other cooperators *only so long as they share the "right" skin color.* But once we admit such a strategy, a niche opens for a new response. Those of the "wrong" skin color can be either racist-blind or racist-savvy. Those who are racist-blind will treat racists as a simple CC, and hence racist-blinds will cooperate with racists. But, since the racist is defecting, the racist-blind will be exploited. Racist-savvy CCs, on the other hand, will detect a racist for what he is: someone who will defect. Hence, the racist-savvy will defect in kind against racists, thus avoiding exploitation.[22] For scoring, I reverse the ordinal ranking, and calibrate it so that the status quo equals "0." That is, defect,coop = 2,-1; coop,coop = 1,1; defect,defect = 0,0; and coop,defect = -1,2.[23] In the exploitation game, then, WP gets 2 points, while BN gets -1. In the non-interaction game, both WP and BN defect against each other. In terms of payoffs for a prisoner's dilemma, both players get 0 points. That is, mutual defection leaves all parties at the status quo, just as in the case of non-interaction.[24]

Game Structure. I have used replicator dynamics in round-robin iterated PD games. Replicator dynamics track successful strategies in terms of number of offspring. If a player using strategy *X* does well in a round, that player has more offspring in the next round, where its offspring play the same strategy. Thus, over time, successful strategies grow in population proportion, while unsuccessful strategies dwindle or go extinct. Thus, successful strategies are ones who take over a larger proportion of the population.[25]

Replicator dynamics in round-robin structures use the following simplifying conveniences: (1) All players play every other player in a single round. (2) Each round represents one distinct generational change, as if every generational

member is born and dies at the exact same instant. (3) Strategies are set; there is no playing the field. Even a mixed strategy is defined prior to the game, and no player playing that mixed strategy can deviate from it. (4) Reproduction is via strict cloning: mutation and recombination common to sexual selection are not part of replicator dynamics. (5) Replicator dynamics measure aggregate population states by the frequency with which certain phenotypes or strategies occur. They have no interest in individual variation within those aggregate populations. The effects of chance fluctuations are deemed to average out.

Not surprisingly, each of these simplifications has met with objection. (1) In our world, we do not interact with every person in our city, let alone our neighborhood. We are more prone to interact with family, friends, associates, acquaintances, and a limited range of business persons. As for the rest, we may not even pass them on the street.[26] (2) Generations overlap in non-uniform patterns. Strategies that might have dwindled can still linger and impact emerging strategies.[27] (3) Real people are rarely so rigid in employing one set strategy for life. Not only might we alter our strategies as we go, but we may also test the field, especially in new contexts.[28] (4) While the replicator dynamics may be well suited to large biological populations tracking sexual selection, they fail to account for cultural evolution produced by imitation and learning. Thus, I need not be stuck with the strategy I was born with (if it even makes sense that I was born with a strategy), but can see the success of my neighbors and alter my own strategy accordingly.[29] (5) Even in a case where the genotype is an exact clone, the phenotype will not necessarily be one. A model allowing for mutation and recombination is more realistic, especially, allowing for variation within groups.[30]

To accommodate the above critiques, a shift has been toward more discrete models. For example, spatialized dynamics account for correlated play by allowing members to interact—not with everyone in staid round-robin games—but within geographical constraints where neighbors play with neighbors, or social networks, where like plays with like.[31] After a specified number of rounds, players swap their strategies for those strategies doing best in that geographic or social niche.[32] Along with these modifications, we can allow "trembling hands" and drift.[33]

I have three reasons for my sticking with the replicator dynamics for examining prejudice. First, my restriction to CC agents already allows the modeling of non-interaction. Mutual defection, in terms of payoffs, is the same as non-interaction. A social network structure with pure agents is the same as permitting mixed strategies in round-robin structures. In social network games, interaction among players is defined by the pre-set network configuration. Players don't play everyone in the game—only those in their social network, perhaps only their immediate geographic neighbors. The network structure can make a difference in whether a strategy evolves in a population or not; just as in nature, a mutant's neighbors will effect whether that mutant passes on its genes or not. The social network games emphasize the importance of local environment. That said, it isn't

the network that makes the difference: it's the neighboring strategies. In a social network game where one interacts with all neighbors (however delineated), a CC will defect against a unilateral defector. But since mutual defection in a PD leaves those mutual defectors at the status quo (no benefit, but no loss either), the result is identical to their not having been on the same social path: that is, non-interaction.[34] That one cooperates with only like-minded sorts and does not interact with non-like-minded sorts is exactly the CC strategy we are starting with. This would not be the case when one is thinking of social neighbors with players not all of whom are of the CC family. Given the PD scoring, my not playing with you in a social network PD yields the same result as my mutually defecting with you in a round-robin PD. Correlated play is simply a kind of mixed play—a strategy that cooperates with certain kinds of individuals but not with certain other individuals. Like plays with like, and ignores the rest.

By non-interaction, I mean in terms of scoring. For a contrast case, consider the Dictator game. *A* offers *B* a division of a piece of pie. Unlike in the Ultimatum game, where *B* has a choice to accept or reject the offer, *B* has *no choice* in the Dictator game but to accept whatever offer *A* has allotted *B*.[35] In such a case, *B* is part of the game, gets a variable score (whatever *A* allots her), yet is totally passive, a "non-interactor." But the sort of "non-interaction" by *B* in the Dictator Game is not the same as "non-interaction" in my games. For me, mutual defection yields the same results as mutual non-interaction. Imagine, for example, Situation 1: *A* passes by a shoe store. There is non-interaction between *A* and the shoe clerk. Now consider Situation 2: *B* enters the shoe store, picks out a pair of shoes, but upon hearing the price for the shoes, withholds her money. Because *B* is withholding her money, the shoe clerk withholds the shoes. In Situation 2, both parties mutually defect from the cooperative bargain of swapping shoes for money. The result is the same as if the customer never entered the shoe store in the first place, Situation 1. In both Situations 1 and 2, all players remain at their status quo: the state of affairs they were in prior to the (pseudo-) interaction. That is, both *A* and *B* have their money, but no shoes, while the shoe clerk has shoes, but no money. Whether we are tracking Situation 1 or 2, therefore, does not matter. Correlated games supposedly allow for Situation 1 as well as 2, but so too—in terms of formal results—do round-robin games with CC agents using replicator dynamics.

Secondly, and more importantly, prejudice may be viewed as a constraint on social networking. What we want to review is how such a constraint would likely arise. A spatialized or social network approach will be unsuited to that task. If the contact thesis is right, people in close contact will tend to lose their prejudices with one another. A game modeling close contact can explain the folly of prejudice, but not the lure of prejudice. If we want more realism in our games, we need a structure that can show the tension between the lure and the folly.

Moreover, replicator dynamics are neutral between biological and cultural

evolution.[36] The use of imitating the best strategy after a number of rounds (using replicator dynamics) may better match the *speed* of cultural evolution, but it is still the phenotypic behavior that is being tracked in the replicator dynamics. Correlated games with imitation, however, rule out the prejudicial maneuver of refusing to imitate the best strategy if the ones employing the best strategy are the wrong color. This is precisely the problem with racism: certain individuals or groups are excluded or exploited for reasons independent of the social strategies they employ. A correlated dynamic seems ill-suited to capture this. Modeling prejudice using replicator dynamics, better than spatialized or social network games, shows that prejudice is both a conditional move that otherwise confers advantage, and—when overextended beyond strategy correlation—maladaptive. It is this tension that explains both the prevalence and the immorality of prejudice.

A third reason to use replicator dynamics instead of other structured games is that Grim and colleagues have already demonstrated the fit of nonprejudice using a spatialized PD with TFTs. They claimed, however, that such results cannot come about using replicator dynamics.[37] So long as we restrict ourselves to CC agents, however, my results show otherwise. That is, my use of replicator dynamics corroborates the findings in Grim and colleagues, despite their doubts.

Game Theory Results

1. *Exploitation Game*. If exploitation is allowed, and equal initial populations prevail, prejudice takes over the population in only four generations. In the case where the initial population proportions are 80% nonprejudice and 20% prejudice, prejudice takes over the population in seven generations.[38] Such results are exactly as one would predict. Basically, a prejudicial agent gets to reap the benefits of both cooperation and defection while never being exploited herself. This is a recipe for success and may well account for the lure of prejudice.

2. *Non-interaction Game*. But prejudice is not always revealed through exploitation. Often it is revealed in simple non-interaction. If a black agent comes into a racist white man's store, the white racist may simply refuse service, not rob the black agent. Moreover, the exploitation game ignores the fact that we are presumably starting from *conditional* cooperation, not *unconditional* cooperation; that is, my games already presume that evolutionary forces have weeded out nonconditional agents. Therefore, it would be bizarre that a BN will tolerate being exploited by a WP. So, even in the case where a WP would want to exploit a BN if possible, a BN should be savvy enough to simply defect in kind. Thus, the simple exploitation game where prejudice wins fails to represent the state of actual practice.

If Ps do not interact with their opposite color, then nonprejudice takes over the population in almost all variations of initial population proportions, typically in about ten generations.[39] The non-interaction game highlights why prejudice

is—to use Jan Narveson's terms—inefficient.[40] As any business student knows, limiting one's customer base is, all things being equal, counterproductive. Also, the non-interaction games merely repeat the lesson evolutionary models have already taught us: returning defection with defection makes defection a non-adaptive trait.

3. *Social Costs Game.* But a world where prejudice is extinct is not the world we currently live in. One way of tinkering with our model until our results are consistent with our real world is to impose social costs for agents who cooperate with out-groups. Social costs for infractions of group norms are certainly common in our world. Following the social cost scale used by Grim and colleagues, I can impose a 0.5 penalty for mixed-color interaction. Starting from equal population proportions, nonprejudice still takes over the population in fifteen generations. If we begin with 80% of the population being prejudiced, nonprejudice takes over the population in twenty-five generations. Here, the end result is nonprejudice, so it appears no more realistic than the non-interaction game, but since the number of generations increases before nonprejudice completely takes over, we can say the results roughly capture our real world state of affairs on the grounds that condemnation of prejudice is slowly (perhaps too slowly) growing, although we are obviously not yet near to having wiped it out.

We could also introduce a mutation rate in the social costs game, to help fix some of the limitations to the replicator dynamics. Introducing mutation rates means that, despite the adaptation of nonprejudice, a continual recurrence of prejudicial traits will extend the time it takes for nonprejudice to completely take over the population, if it is ever possible at all. So, especially with a mutation rate, the social costs game is an improvement in terms of realistic modeling. Despite social costs for nonprejudice, nonprejudice still prevails in the long haul.

4. *Boycotter Game.* Applying social costs is admittedly ad hoc. Presumably, the reason one group would form a particular cultural norm (or meme) is to preserve an adaptive trait. We do not say stealing is bad because one will be punished for it. Rather, we suppose we punish stealing because stealing is bad for us. Similarly, if we apply a social cost for consorting with an out-group, it is because there is something maladaptive in mingling with that out-group independent of the social punishment we invoke. The social cost method fails to account for that. Another method to stem prejudice is to boycott interactions with the prejudiced. The dissolution of apartheid, for example, is in large part credited with international embargoes against South Africa. To be a boycotter, then, is not merely to not cooperate with those who do not cooperate with you, but to not cooperate with noncooperators, period—even if they would have cooperated with you. Boycotters, then, can be seen as moral crusaders, taking the costs upon themselves in order to punish racists.

As expected, prejudicial dispositions fare poorly if boycotters exist and, generally, in less than ten generations. But if our goal is to account for the rise and fall of prejudice by purely evolutionary forces, any new dispositional trait must itself

confer fitness to the carrier of that trait. That is, although boycotters undermine the success of prejudice, they cannot be allowed into the game if another disposition does better. Except for a few cases, while boycotters do undermine prejudice, boycotters do not do better, and often do worse, than nonprejudice. As a result, we cannot imagine that a niche for boycotters exists. For example, boycotters do better than the merely nonprejudiced only if either boycotters get to exploit the prejudiced (not likely), or so long as the initial population proportion favored boycotters over the merely nonprejudiced—again unlikely, evolutionarily speaking.

Moreover, the complaint against introducing social costs applies equally well with boycotter solutions. Boycotting is an intentional act based on moral reasons. But our role is to account for the origin of those moral reasons. Introducing boycotters begs that question.

5. *Probability Game*. So far, this is what we have learned: Prejudice has evolutionary advantage so long as the circumstances are closer to the exploitation game, and our moral outrage against prejudice takes root so long as the circumstances are closer aligned with the non-interaction game, and that the reality of our current world seems to be somewhere in between. That is, some of the time prejudicial behaviors may well be exploitive, and at other times, merely incur non-interaction. To account for this mixture in game theoretic terms without imputing ad hoc costs, or introducing new dispositions, we can simply vary the probability of exploitation.

If prejudicial strategies exploit members of the out-group 20% of the time, that is still not sufficient to make prejudice a viable strategy. A 20% exploitation rate merely extends the number of generations it takes for nonprejudice to prevail, but prevail they do—no matter the initial population proportions (eighteen generations from equal initial population proportions, twenty-six generations when 80% of the population are prejudiced to begin with). A 40% exploitation rate, however, turns the tides in favor of prejudice. From equal population proportions, prejudice prevails after nineteen generations. So, a 20% exploitation rate is too low, while 40% is too high. A 30% exploitation rate, on the other hand, alters things in favor of prejudice only so long as the prejudiced outnumber the nonprejudiced in the initial population. Notice, particularly, the middle two Start/Finish columns of Table 1. Here, an interesting polymorphism prevails from initial populations where the prejudiced outnumber the nonprejudiced 3:2.

Table 1. 30% exploitation rate. Numbers in parentheses indicate number of generations until equilibrium is reached (with rounding).[41]

	Start	Finish (32)	Start	Finish (3)	Start	Finish (14)
WN	0.25	0.5	0.2	0.16	0.1	0
WP	0.25	0	0.3	0.34	0.4	0.5
BN	0.25	0.5	0.3	0.16	0.4	0
BP	0.25	0	0.2	0.34	0.1	0.5

Assuming that the exploitation rate will fluctuate over time, as opposed to remaining stable, it is not far-fetched to say these preliminary findings support the lure of prejudice despite the opposition to prejudice. Thus, we can explain the prevalence of a trait we deem immoral while at the same time claiming that our very deeming it immoral is an evolutionary adaptation.

Conclusion

By focusing purely on naturalistic dynamics, we can explain both the lure of prejudice and the rise of our moral indignation to prejudice. Prejudice is not an adaptation. Rather, it piggybacks on conditional cooperation, which is an adaptive trait. Part of this piggybacking is due to fluctuations in heuristics, part of this is due to fluctuations in population makeup, and part of this may be explained by the mixed-signal hypothesis. In all cases, something like the contact thesis for reducing prejudice is corroborated.

University of Prince Edward Island

NOTES

Thanks to Esther Kemp and Bill Harms for computer programming help on the simulations. Thanks also to Paul Viminitz, Scott Woodcock, and two anonymous reviewers for valuable commentary.

1. We do not think it odd to hear how many Americans were killed in Afghanistan, but the tragedy befalls anyone killed, not just Americans. As Basu notes, imagine the uproar if the news reported the number of whites killed in Afghanistan today, or the number of Catholics killed in Afghanistan today. Kaushik Basu, "Racial Conflict and Malignancy of Identity," *Journal of Economic Inequality*, vol. 3, no. 3 (2005), pp. 221–241, particularly p. 223.

2. See, for example, Pat Shipman, *The Evolution of Racism* (Cambridge, MA: Harvard University Press, 2002).

3. John Maynard Smith, "Science and Myth," in *The Philosophy of Biology*, eds. David Hull and Michael Ruse (Oxford: Oxford University Press, 1998), p. 374.

4. "Like-minded" is best translated as "similarly programmed" to avoid assuming our agents are psychological beings. This worry does not come up for "mind" reductionists.

5. For the success of TFT in Prisoner Dilemma (PD) games, see Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984); and James W. Friedman, "A Non-Cooperative Equilibrium for Supergames," *Review of Economic Studies*, vol. 38, no. 113 (1971), pp. 1–12. For reciprocal altruists, see Robert Trivers, "The Evolution of Reciprocal Cooperation," *Quarterly Review of Biology*, vol. 46, no. 1 (1971), pp. 35–57.

For constrained cooperation in PDs, see David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986); and Peter Danielson, *Artificial Morality: Virtuous Robots for Virtual Games* (London: Routledge, 1992). For the "fairman" strategy in ultimatum games, see Brian Skyrms, *The Evolution of the Social Contract* (Cambridge: Cambridge University Press, 1996). For queuing behaviors in chicken-like games, see Malcolm Murray, *The Moral Wager* (Dordrecht, Netherlands: Springer, 2007), pp. 117–121.

6. Conditional cooperation is a strategy. A conditional cooperator is a player who adopts the CC strategy. For convenience, I label both the strategy and the player "CC." In context, this should be clear. For other strategies that belong to the same family as CC, see Robert Sugden, *Economics of Rights, Cooperation and Welfare* (New York: John Wiley and Sons, 1986); Michihiro Kandori, "Social Norms and Community Enforcement," *Review of Economic Studies*, vol. 59, no. 1 (1992), pp. 63–80; and Glenn Ellison, "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching," *Review of Economic Studies*, vol. 61, no. 3 (1994), pp. 567–588.

7. Apart from Axelrod, Danielson, Ellison, Gauthier, Kandori, Skyrms, and Sugden, see J. McKenzie Alexander, "Group Dynamics in the State of Nature," *Erkenntnis,* vol. 55, no. 2 (2001), pp. 169–182; and Peter Danielson, "Evolutionary Models of Co-Operative Mechanisms: Artificial Morality and Genetic Programming," in *Modeling Rationality, Morality, and Evolution*, ed. Peter Danielson (Oxford: Oxford University Press, 1989), pp. 423–441.

8. Defection is the rational choice in a single PD, since whatever the other player does, you do one better defecting compared to cooperating. As Binmore reminds us, "[i]f the players have the power to alter their preferences or to commit themselves to behaving in ways before the play of the Prisoner's Dilemma, then it is not the Prisoner's Dilemma that they are playing." Kenneth G. Binmore, *Game Theory and Social Contract: Vol. 1: Playing Fair* (Cambridge, MA: MIT Press, 1998), p. 27. There are situations in which the prudent resolve under uncertainty may be to cooperate, however. This would be when the odds of bumping into a cooperator are sufficiently higher than the odds of bumping into a defector. Such a world is not Hobbes's state of nature. If we are no longer in a state of nature, and cooperators sufficiently outnumber defectors, and one's detection skills are only periodically damaged, then testing the waters with a tentative "coop" move when in doubt will do better than the so-called prudent "defect." But part of the lure of replicator dynamics is to explain how we got to such a state in the first place. Similarly, one might be inclined to conceive a TFT as a strategy in which "cooperate" is the default. Certainly this is so on the first move. Since TFT generally prevails in competitions, this may undermine treating "defect" as the prudent default. But here, I am speaking of uncertainty. To make that fit the TFT strategy, a TFT agent will be unable to remember a previous play. A memory-taxed TFT would be indistinguishable from an unconditional cooperator, and would fare poorly. A partially forgetful TFT, on the other hand, may be closer in line to Tit-for-Two-Tats.

9. Paul Viminitz has made a case for this in his game theoretic analysis of racism (unpublished).

10. A TFT responds to you according to how you responded with him in the previous play. A CC, on the other hand, predicts how you will likely respond in the current interaction. Thus, while a TFT can get exploited by a defector on the first play, a CC should avoid that initial cost. Thus, a CC can do better than a TFT.

11. James Baldwin, "Sonny's Blues," in *Going to Meet the Man* (New York: Dial Press, 1965), p. 118.

12. Gordon Willard Allport, *The Nature of Prejudice* (Cambridge, MA: Addison-Wesley, 1954).

13. See Basu, "Racial Conflict," p. 224, who cites J. C. Li, David Dunning, and Roy S. Malpass, "Cross-Racial Identification among European Americans: Basketball Fandom and the Contact Hypothesis," mimeo (University of Texas, El Paso, 1998).

14. Patrick Grim, Evan Selinger, William Braynen, et al., "Modeling Prejudice Reduction: Spatialized Game Theory and the Contact Hypothesis, *Public Affairs Quarterly*, vol. 19, no. 2 (2005), pp. 95–125. I use CC agents in nonspatialized games and, counter to Grim et al.'s expectation, get similar results.

15. Peter Kollock, "Transforming Social Dilemmas," in *Modeling Rationality, Morality, and Evolution*, ed. Peter Danielson, pp. 185–209. See also Morton Deutsch, "Social Psychology's Contributions to the Study of Conflict Resolution," *Negotiation Journal*, vol. 18, no. 4 (2002), pp. 307–320.

16. For Binmore's objection to Social Darwinism, see Binmore, *Game Theory*, pp. 99–100. See also Richard Joyce, *The Evolution of Morality* (Cambridge, MA: MIT Press, 2007), pp. 221–222.

17. Stephen Jay Gould, *The Panda's Thumb: More Reflections in Natural History* (Harmondsworth, England: Penguin, 1980).

18. Brian Garvey gives the analogy of tinkering with your car engine. Without knowing what you are doing, a small tinker might make the car work better, whereas a large tinker will much more likely make the engine stop working altogether. Brian Garvey, *Philosophy of Biology* (Montreal: McGill-Queen's, 2007), p. 11.

19. See, for example, Stephen Jay Gould and Elisabeth Vrba, "Exaptation: A Missing Term in the Science of Form," *Paleobiology*, vol. 8, no. 1 (1982), pp. 4–15; or Stephen Jay Gould, "Exaptation: A Crucial Tool for Evolutionary Psychology," *Journal of Social Issues*, vol. 47, no. 3 (1991), pp. 43–65. See also John H. Ostrom, "Archaeopteryx and the Origin of Flight," *Quarterly Review of Biology*, vol. 49, no. 1 (1974), pp. 27–47; or John H. Ostrom, "Bird Flight: How Did It Begin?" *American Scientist*, vol. 67, no. 1 (1979), pp. 46–56.

20. See Reinhard Selten, "What Is Bounded Rationality?" in *Bounded Rationality: The Adaptive Toolbox*, eds. Gerd Gigerenzer and Reinhard Selten (Cambridge, MA: MIT Press, 2002), pp. 13–36; or Robert Boyd and Peter Richerson, *Culture and the Evolutionary Process* (Chicago: University of Chicago Press, 1985).

21. James L. Gould, *Ethology: The Mechanism and Evolution of Behavior* (New York: Norton, 1982), p. 164.

22. I oversimplify in the text. I mean, for example, a black racist-savvy will defect against a white racist. Whether a black racist-savvy will cooperate with a black racist, for example, is a separate issue—one I take up in the boycotter game.

23. Calibrating mutual defection at 0,0 highlights that mutual defection is the status quo. When scoring is {4,1; 3,3; 2,2; 1,4}, the status quo is 2,2. Getting 2 points is more than getting 0 points by not playing at all. This is a contaminant of reverse ordinal ranking

without calibrating to zero. Otherwise, there is no effect on the dynamical processes by the altered payoffs.

24. At least this is true for PD, *Chicken*, and *Battle of the Sexes*. In *Chicken*, simultaneous swerving produces the same payoff as not having played at all. In *Battle of the Sexes*, if She prefers to go to the movies, while He prefers to go the opera, a date where She goes to the movies and He goes to the opera is the status quo—the payoff equal to not having played at all. In the *Ultimatum Game*, things are not quite so simple. Not playing the game at all entails zero pieces of pie. Only certain pairings of strategies can equal that state (two reject-alls, for example). Likewise in zero-sum games, the option of remaining at one's status quo is available only in draws, and only in those kinds of games or game reiterations where draws are possible and the utility of playing to a draw is not itself positive.

25. See Peter D. Taylor, and Leo B. Jonker, "Evolutionary Stable Strategies and Game Dynamics," *Mathematical Biosciences,* vol. 40, no. 1–2 (1978), pp. 145–156; Jörgen W. Weibull, *Evolutionary Game Theory* (Cambridge, MA: MIT Press, 1998), pp. 69–119; or Avinash K. Dixit, Susan Skeath, and David H. Reiley, Jr., *Games of Strategy* (3rd edition) (New York: W. W. Norton, 2009); and Larry Samuelson, *Evolutionary Games and Equilibrium Selection* (Cambridge, MA: MIT Press, 1997).

26. See, for example, J. McKenzie Alexander, *The Structural Evolution of Morality* (Cambridge: Cambridge University Press, 2007), p. 26.

27. See Weibull, *Evolutionary Game Theory*, pp. 124–126.

28. See Michihiro Kandori, George J. Mailath, and Rafael Rob, "Learning, Mutation and Long-Run Equilibria in Games," *Econometrica*, vol. 61, no. 1 (1993), pp. 29–56; and H. Peyton Young, *Individual Strategy and Social Structure* (Princeton, NJ: Princeton University Press, 1998).

29. See Boyd and Richerson, *Culture and the Evolutionary Process*; Dan Sperba, *Explaining Culture: A Naturalistic Approach* (Oxford: Oxford University Press, 1996); and Samuelson, *Evolutionary Games and Equilibrium Selection.*

30. John Gale, Kenneth G. Binmore, and Larry Samuelson, "Learning to Be Imperfect: The Ultimatum Game," *Games and Economic Behavior*, vol. 8, no. 1 (1995), pp. 56–90; and Dean P. Foster and H. Peyton Young, "Stochastic Evolutionary Game Dynamics," *Theoretical Population Biology*, vol. 38, no. 2 (1990), pp. 219–232. But see also Alexander, *The Structural Evolution of Morality*, p. 37. For an account of the structural stability of competing models, see Brian Skyrms, "Stability and Explanatory Significance of Some Simple Evolutionary Models," *Philosophy of Science*, vol. 67, no. 1 (2000), pp. 94–113.

31. Robert Aumann, "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, vol. 1, no. 1 (1974), pp. 67–96; Robert Aumann, "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, vol. 55, no. 1 (1981), pp. 1–18; Robert Boyd and Peter J. Richerson, "Norms and Bounded Rationality," in *Bounded Rationality: The Adaptive Toolbox*, eds. Gerd Gigerenzer and Reinhard Selten (Cambridge, MA: MIT Press, 2002), pp. 281–296; and Jason Alexander and Bryan Skyrms, "Bargaining with Neighbors: Is Justice Contagious?" *Journal of Philosophy*, vol. 96, no. 11 (1999), pp. 588–598. For variations on dynamic networks, see Alexander, *The Structural Evolution of Morality*, pp. 38–52.

32. See, for example, Martin A. Nowak and Robert M. May, "Evolutionary Games and Spatial Chaos," *Nature*, vol. 359, no. 6398 (1992), pp. 826–829; Kristian Lindgren and Mats G. Nordahl, "Evolutionary Dynamics of Spatial Games," *Physica D*, vol. 75, no. 1–3 (1994), pp. 292–309; Joshua M. Epstein, "Zones of Cooperation in Demographic Prisoner's Dilemma," *Complexity*, vol. 4, no. 2 (1998), pp. 36–48; and Brian Skyrms, *The Stag Hunt and the Evolution of Social Structure* (Cambridge: Cambridge University Press, 2004), pp. 23–29.

33. "Trembling hand" allows for players to periodically respond in ways counter to their built-in strategies. Drift allows random strategies to periodically stumble into the game beyond the ones replicator dynamics predict. See Reinhard Selten, "Re-Examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, vol. 4, no. 1 (1975), pp. 22–55. See also Samuelson, *Evolutionary Games and Equilibrium Selection*, p. 51; and Weibull, *Evolutionary Game Theory*, pp. 20–21.

34. This assumes the PD captures bargaining interactions. If we are speaking of warfare, mutual defection can leave both parties worse off than their status quo, for example, dead.

35. In psychological experiments with real people (as opposed to computer simulations), fair divisions still tend to prevail. Unfair divisions prevail only when *A* feels her division will not be traced back to her personally. See Dixit, Skeath, and Reiley, Jr., *Games of Strategy*, pp. 72–73.

36. Alexander, *Structural Evolution of Morality*, p. 28.

37. Grim et al., "Modeling Prejudice Reduction," pp. 96–97.

38. Notice that I alter the initial population proportion in favor of the loser in this game, something I do throughout. Except for the probability game discussed later, I use a 50–50 and 80–20 population mix between prejudicial and nonprejudicial attitudes to show that altering initial population mixes only delays the number of generations it takes to reach equilibrium; it does not alter the eventual equilibrium. In the case of the probability game, however, things are different. In that game, altering the population proportion can alter the eventual equilibrium outcome, and so I offer a more refined analysis (see Table 1).

39. Given a score of 0,0 for mutual defection, whether we model mutual defection or simple non-interaction will not matter as far as results go.

40. Jan Narveson, *The Libertarian Idea* (Philadelphia: Temple University Press, 1988), pp. 183–184.

41. To get these results, we begin with the basic PD payoffs using reverse ordinal ranking calibrated at 0 for status quo. Thus,

	coop	defect
coop	1,1	–1,2
defect	2,-1	0,0

Using this guide, the payoffs for the various strategies on a single round where Ps get to exploit Ns 30% of the time is so,

	WN	WP	BN	BP	average
WN	1	1	1	0(.7) –1(.3)	0.675
WP	1	1	0(.7) +2(.3)	0	0.65
BN	1	0(.7) –1(.3)	1	1	0.675
BP	0(.7) +2(.3)	0	1	1	0.65

To chart the replicator dynamics for strategy *A*, we use the formula: U(*a*)p(*a*)/U. U(*a*) = the average utility for a player using strategy *A*. p(*A*) = the current population proportion of players using strategy *A*. U = the average utility of all players.