"Machiavellian" Intelligence as a Basis for the Evolution of Cooperative Dispositions

JOHN ORBELL *University of Oregon*
TOMONORI MORIKAWA *Waseda University*
JASON HARTWIG *University of Oregon*
JAMES HANLEY *University of Illinois, Springfield*
NICHOLAS ALLEN *University of Melbourne*

How to promote cooperative behavior is classically solved by incentives that lead self-interested individuals in socially desirable directions, but by now well-established laboratory results show that people often do act cooperatively, even at significant cost to themselves. These results suggest that cooperative dispositions *might be an evolved part of human nature. Yet such dispositions appear inconsistent with the "Machiavellian intelligence" paradigm, which develops the idea that our brains have evolved, in substantial part, for capturing adaptive advantage from within-group competition. We use simulation to address the evolutionary relationship between basic Machiavellian capacities and cooperative dispositions. Results show that selection on such capacities can (1) permit the spread of cooperative dispositions even in cooperation-unfriendly worlds and (2) support transitions to populations with high mean cooperative dispositions. We distinguish between "rationality in action" and "rationality in design"—the adaptive fit between a design attribute of an animal and its environment. The combination of well-developed Machiavellian intelligence, modest mistrust,* and *high cooperative dispositions appears to be a rational design for the brains of highly political animals such as ourselves.*

How to promote group-benefiting action when individuals have strong incentives against such action is perhaps the fundamental problem of political theory (Axelrod 1981; Ostrom 1998). Classically spelled out by Hobbes ([1651] 1947) but now usually framed in terms of the Prisoner's Dilemma, the issue is how to persuade individuals to act in the interests of the collectivity (to "cooperate") when there is a clear incentive for them to do otherwise (to "defect"). The incentive structure means that there is only one solution for rational and self-interested individuals: The incentives they confront must, somehow, be changed so that cooperation rather than defection offers the greater return. We now understand that this does not necessarily imply a centralized Hobbesian Leviathan, and decentralized mechanisms might be sufficient—for example, by the existence of "altruistic punishment" (Boyd et al. 2003). Nevertheless, without additional or "selective" (Olson 1965) incentives, rational and self-interested individuals *will* defect in Prisoner's Dilemma situations.

But that does not mean that real people confronting real PD-like situations will always defect. In fact, as Field (2001) has recently emphasized, there is now strong laboratory evidence documenting humans' frequent willingness to cooperate even in one-shot Prisoner's Dilemmas where the incentive to defect is substantial and unambiguous (e.g., Caporael et al. 1989; Orbell and Dawes 1993; and Ostrom, Walker, and Gardner 1992). The incidence of such cooperation varies between studies and experimental conditions, but people often *do* cooperate in prisoner's dilemmas, sometimes in large numbers. The extensive literature on ultimatum and dictator games, although not addressing cooperation per se, supports the same general conclusion: People *are* frequently prepared to carry significant private costs to benefit of others (Camerer and Thaler 1995).

Taking such findings seriously has implications for how we address classic problems in political science, notably normative issues surrounding the design of institutions. Certainly, it is foolish to design institutions in such a way that socially desirable outcomes depend on people consistently acting *against* their private interests (G. Hardin 1977). But "incentive compatibility" and cooperative dispositions might interact. On the positive side, perhaps awareness that "compatible" incentives are in place will increase players' expectations that others will cooperate, increasing thereby the likelihood of their own cooperative dispositions being expressed in their own behavior. This might happen, for example,

John Orbell is Professor, Department of Political Science, 936 PLC Hall, 1415 Kincaid Street, University of Oregon, Eugene, OR 97403-1284 (jorbell@uoregon.edu).
Tomonori Morikawa is Professor, International College, Waseda University, Tokyo, Japan (goducks@waseda.jp).
Jason Hartwig is Instructor, Political Science Department, University of Oregon, Eugene, OR 97403. (jhartwig@uoregon.edu).
James Hanley is Assistant Professor, Political Studies Program, University of Illinois, Springfield, IL 62703 (jh2@hanley.net).
Nicholas Allen is Principal Research Fellow, ORYGEN Research Centre, and Associate Professor, Department of Psychology, University of Melbourne, Melbourne, Australia (n.allen@psych.unimelb.edu.au).

Oleg Smirnov gave us valuable critical and statistical help, and Ron Mitchell provided constructive readings of early drafts. Cian Montgomery helped develop an earlier simulation out of which the present one grew. The College of Arts and Sciences at the University of Oregon gave critical financial assistance in the early stage of the project, which was subsequently funded by the National Science Foundation Grant 9808041. Morikawa is partly funded by a 2003 Grant-in-Aid for Scientific Research of the Japan Society for the promotion of Science, under the title The Evolution of Political Intelligence and Decision-Making.

when responses are governed by fear of being suckered rather than by greed for the free rider's payoff (Orbell et al. 1986). On the negative side, Frohlich and Oppenheimer (1996) have shown that incentive compatibility can, sometimes at least, undermine socially productive behavior that would otherwise result from innate cooperative dispositions. Perhaps the social capital represented by such dispositions would be substantially lost if we were to organize social life exclusively as a response to private incentives. In general, the now well-documented existence of innate cooperative dispositions means that it is almost certainly an error to base recommendations about institutional design exclusively on "incentive compatibility"—or on the assumption that cooperative dispositions are universal and infallible in their production of cooperative behavior.

We do take those findings seriously and believe, with Field (2001), that it is appropriate to address the problem in biological—that is to say, evolutionary—terms. Doing so, of course, goes against the "standard social science model" (Tooby and Cosmides 1992) that seeks explanations of human behavior exclusively in cultural or environmental terms, but drawing a strict dichotomy between "culture" and "biology" is now widely recognized as erroneous (e.g., Boyd and Richerson 1985 and Dunbar, Knight, and Power 1999). Humans are certainly cultural animals, and culture certainly has a lot to do with human behavior, including cooperation. But accepting these facts only poses further evolutionary questions, for example, How and why have humans evolved the capacity for culture? What are the constraints that evolved human nature places on the substance of culture? and What are the constraints that evolving culture places on human nature? Granted that the issue is human *nature*, biology is the ultimate source of understanding—and, as Dobzhansky (1973) famously pointed out, "Nothing in biology makes sense except in the light of evolution."

Several evolutionary arguments about cooperative dispositions are well known in the life sciences. The theory of kin selection (Hamilton 1964) shows how a disposition to assist others with whom one is related can rebound to one's own *genetic* advantage even if adaptive costs are involved in the helping act, suggesting that our cooperative responses toward nonkin might be founded, at least in part, on an evolved disposition to cooperate toward kin. The theory of reciprocity (Axelrod 1984; Trivers 1971) provides an "I'll scratch your back if you scratch mine" explanation for cooperative dispositions that does not require genetic relatedness, but only that individuals (even members of different species) encounter each other through an iterated sequence of Prisoner's Dilemma-like games. A further model, the one favored by Field (2001), is group selection, now staging a comeback (Sober and Wilson 1998) after many years of having been widely discounted (Dawkins 1976; Williams 1966). Although still hotly disputed (Reeve 2000), group selection's capacity to promote cooperative dispositions requires that a group's survival prospects be increased by members' cooperative choices, with the cooperator's fitness gains from the group's success being greater than the fitness costs that individuals incur as a result of their cooperative choices.[1]

Notably absent from this list is any model of how cognitive capacities designed by natural selection for addressing prior problems might have provided a sufficient, even necessary, basis from which cooperative dispositions could subsequently evolve. As is widely recognized, natural selection must build from existing structures—in Dennett's (1995) terms, using "cranes" rather than "skyhooks"—and any significant cooperative dispositions must have evolved in the context of prior adaptations. In this spirit, the broad question we address is, *Might cognitive capacities originally designed to address other adaptive problems have made possible the evolution of cooperative dispositions such as those now strongly suggested by the empirical data*?

In Tinbergen's (1963) famous distinction, our interest is in "ultimate" rather than "proximate" processes—that is, we address adaptive pressures across many generations as opposed to particular mechanisms that might have evolved to capture adaptive gains in the here-and-now. This does not detract, of course, from the interest attaching to proximate mechanisms. At the conclusion of his original paper, for example, Trivers (1971) laid out a formidable research agenda for psychology by proposing that moralistic aggression, gratitude, sympathy, guilt, friendship, gossip, cheating, and the ability to detect cheating might have evolved as proximate mechanisms for capturing the gains available from reciprocity. But in evolutionary theory, ultimate selective pressures and proximate mechanisms are distinct issues, and we do not address the latter here.

Our particular interest is in cognitive mechanisms fundamental to the "Machiavellian intelligence" (Byrne and Whiten 1988; Whiten and Byrne 1997) hypothesis.[2] In its broadest terms, this proposes that group living selects strongly for whatever cognitive capacities facilitate an individual's successful negotiation of the competitive and highly complex social environment of the group. A seminal paper by Humphrey (1976; see also Jolly 1966) pointed to the complexity, fluidity, and recursiveness of social relations within primate groups as posing a particularly difficult adaptive problem for members of those groups—far more difficult, Humphrey argued, than the standard adaptive problems of gathering resources and avoiding predators:

> Once a society has reached a certain level of complexity, then new internal pressures must arise which act to increase its complexity still further. For . . . an animal's 'adversaries' are members of his own breeding community. If intellectual

[1] We use "fitness" in the standard (although sometimes debated) evolutionary sense of an individual's relative success in populating subsequent generations with its descendants. Adaptations—complex structures produced by natural selection in response to challenges in a species' ancestral "environment of evolutionary adaptation"(Bowlby 1969)—necessarily involve fitness costs as well as fitness benefits, with the presumption being that an adaptation would not have been selected for if the net were not positive.

[2] Known also as the "social" (Brothers 1997) and "political" (Boehm 1997) intelligence hypothesis.

> prowess is correlated with social success, and if social success means high biological fitness, then any heritable trait which increases the ability of an individual to outwit his fellows will soon spread through the gene pool. And in these circumstances there can be no going back: an evolutionary 'ratchet' has been set up, acting like a self-winding watch to increase the general intellectual standing of the species. (311)

Although there are substantial adaptive advantages to group living—most obviously, better protection from predators, improved success *as* predators, and more ready access to mates—there is also genetic competition between group members, meaning that those attributes best equipped to win the never-ending Machiavellian games of "social chess" or "plot and counter plot" will be positively selected.

> Thus social primates are required by the very nature of the system they create and maintain to be calculating beings; they must be able to calculate the consequences of their own behavior, to calculate the likely behavior of others, to calculate the balance of advantage and loss—and this all in a context where the evidence on which their calculations are based is ephemeral, ambiguous and liable to change, not least as a consequence of their own actions. (309)

By this hypothesis, evolutionary competition between group members selects for capacities that allow individuals to deceive and exploit other group members, while simultaneously avoiding being deceived and exploited by others. Employing the terms introduced by Dawkins and Krebs (1978), the two fundamental "Machiavellian" capacities are (1) Sender's capacity to persuade another group member to accept as true what it is in Sender's interest to have it believe is true—viz, *manipulation*[3]—and (2) Receiver's capacity to penetrate to the truth underlying messages from potentially manipulative others—viz, *mindreading*. More specifically, therefore, our research question is, *Could evolved "Machiavellian" capacities for manipulation and mindreading be a basis for the evolution of cooperative dispositions among social animals such as ourselves*?

This has the air of paradox. On the one hand, the Machiavellian intelligence hypothesis seems to dovetail quite nicely with the "harder" models of rational action, implying a brain designed to facilitate the pursuit of private welfare by whatever means necessary. On the other, cooperative behavior involves, by definition, rejecting a dominant incentive, an alternative that is superior for the acting individual regardless of what another individual might choose. If our brains are designed for "Machiavellian" purposes, should we not expect evolved dispositions to *defect*—at least when we can get away with it—not to cooperate?

We use evolutionary simulation to show how, under specified parameters, it is not only possible that Machiavellian capacities provide an evolutionary foundation for cooperative dispositions, but in fact quite likely. If we are, by nature, Machiavellian animals benefiting from group living but also exploiting other group members when that is possible, our findings show how Machiavellian capacities could be a critical underpinning for evolutionary selection on cooperative dispositions capable of motivating us, perhaps quite often, to act cooperatively—thus helping to resolve collective action problems *absent* appropriate selective incentives.

DESIGN OF THE SIMULATION

We use simulation for two reasons. First and most obviously, the processes that interest us occurred over a period of perhaps hundreds of thousands of years and, because they involve social relationships, could leave only little by way of a directly observable fossil record (Wynn 2002). Although careful inference from the existing fossil record can help us understand when particular cognitive structures evolved (e.g., Mithen 1996), such work is necessarily informed by plausible models of process, and simulation is one way of constructing such models. Second, simulation can be a tool for discovery, letting the researcher explore the consequences of diverse parameter settings and, perhaps, identify theoretically interesting processes that might not have been considered using standard analytic techniques. This has been our approach—meaning that a first order of business here is to specify the design decisions made in constructing the simulation.

A simulation for studying the relationship between Machiavellian capacities and cooperative dispositions requires modeling decisions at several levels of analysis:

Individual attributes: We incorporate two dispositions and two cognitive (information processing) capacities. The dispositions are, first, to act in a cooperative manner within joined Prisoner's Dilemma games[4] and, second, to "mistrust" others' willingness to do so (thus its converse, to "trust" them). The cognitive capacities are, first, the capacity to manipulate the persuasiveness of what the individual communicates—true or false—to others and, second, the capacity to mindread the truth underlying others' efforts at such manipulation (Dawkins and Krebs 1978). At the ultimate or functional level at which we operate, of course, we can sidestep the always fascinating proximate questions about particular mechanisms facilitating such manipulation and mindreading.

Interindividual communication: Because we are interested in the evolution of cooperation, when two

[3] As Trivers (1985) has commented, "One of the most important things to realize about systems of animal communication is that they are not systems for the dissemination of the truth. An animal selected to signal to another animal may be selected to convey correct information, misinformation, or both" (395).

[4] Although much recent theoretical and empirical work has focused on cooperation within multiple-person Prisoner's Dilemma games, the problem of predicting others' behavior in such games is computationally formidable, perhaps requiring simplifying heuristics of one kind or another (Orbell and Dawes 1993), whereas the problem of choice and incentives remains basically the same in two-person and n-person PDs (although see Dawes 1975). Accordingly, we focus exclusively on the two-person PD here—defined, as usual, by a game with (1) a binary choice between "cooperation" and "defection"; (2) the standard payoff notation being t = lone defection or free riding, c = mutual cooperation, d = mutual defection, and s = lone cooperation or "being suckered"; and (3) $t > c > d > s$ and $2c > (t + s) > 2d$.

individuals encounter each other they must choose between entering and not entering a potentially cooperative game. In this context, Machiavellian intelligence requires that each sends messages to the other about his or her intentions within any such game (subject to manipulation) and that each evaluates the truth-value of such messages received from the other (a task performed by mindreading).

Individual decision-making: Individuals must make "social" decisions—whether to enter a particular potentially cooperative game and, if they do, whether to cooperate or defect—and those decisions must be informed, at least in part, by their estimates of what the other individual is likely to do. Importantly, the option of *not* entering such a game implies the availability of an alternative course of action that has some value, thus that the individual's play vs. not play decision will be based on its comparison between what it expects to get from playing a cooperative game and what it expects from such an alternative. There are, of course, many different games that social animals play—Rapoport et al. (1976) identified 78 logically distinct two-by-two games, and many of those have multiple-person variants—but modeling a social ecology with all that complexity, if indeed possible, would detract from our primary goal of understanding how the evolution of Machiavellian cognitive capacities and cooperative dispositions might be related, although, particularly interesting alternatives are (1) making a "hawk" challenge, indicating a willingness to fight over some resources otherwise in control of another individual (Orbell, Morikawa, and Allen 2002), and (2) acting in a solitary, "go it alone" manner, extracting resources from the environment absent any interaction with one's own species—an entirely "nonsocial" course of action.

Natural selection: In an evolutionary model there must be a process by which individuals' actions during their lifetimes are translated into relative reproductive success, thus that allows for selective retention in the population of whatever attributes produced that success (Dennett 1995). As is well recognized, population attributes can change by *drift* as well as by relative reproductive success; in fact, we will show that drift might have played an important role in the evolution of cooperative dispositions. In either case, however, evolution requires a source of *variation* that is random with respect to the individual's success but that makes it possible for some individuals to be reproductively more successful than others. We recognize that variation in the natural world can be produced by, for example, recombination as well as by mutation, but for simplicity in what follows we use the term "mutation" to include all sources of variation. Critically for our purposes, mutations must happen on the dispositional and cognitive attributes that bear on an individual's success in complex social environments.

Dispositional and Cognitive Attributes

We model an individual's disposition to cooperate as a probability between 0.0 and 1.0 (its probability of cooperating, or PC). Each decision it makes is determined by a random draw from its PC; it will cooperate if the draw falls at or below its PC and defect if it falls above that value. Notice that such decisions are not appropriately thought of as "rational." Granted, an individual with a low PC might be classified as more "rational" than one with a higher such value; cooperation is, after all, dominated. But that has no particular bearing on our evolutionary model in which such an individual is simply "undisposed to cooperate." If such dispositions are adaptive, then they will be positively selected and will spread throughout the population. If they are not, then they will be selected against and, probably, will vanish—"rational" or not.

We assume that every agent, on encountering another, makes the claim "I will always cooperate" as it makes no strategic sense—certainly in "Machiavellian" terms—to claim anything less than perfect cooperativeness. The total communication is modeled as 100 messages, some of which ($PC * 100$) are true (reflecting the true probability of cooperating) and the remainder of which ($\{1 - PC\} * 100$) are false (reflecting the true probability of noncooperation). But the sender of any such communication confronts the discrete manipulative problems of being convincing in both its true and its false messages. Each individual is therefore equipped with separate manipulative capacities for its true messages (Manipulate True) and for its false ones (Manipulate False), in each case varying between 0.0 and 1.0. Agents in our simulation might, therefore, be good truth-tellers *and* good liars, good truth-tellers but poor liars—and so on. Particular truths and particular lies are not all equally persuasive but are normally distributed around the agent's mean manipulative capacity in each case.

Correspondingly, individuals are modeled as having a mindreading capacity that varies between 0.0 and 1.0. At the low extreme of mindreading, an individual has *no* capacity to penetrate the truth of a potential partner's "I will always cooperate" claim. While at the high extreme it has a perfect capacity to do so, being (potentially) able to recognize all Sender's lies as lies and all its truths as truths. Notice that we do not model mindreading capacities specialized separately for responding to truths and lies. That would imply a cognitive apparatus equipped with *a priori* knowledge of which is which, thus that the problem of mindreading has already been solved.[5]

"Trust" is a concept that has attracted much interest in recent years (e.g., Fukuyama 1995 and Nesse 2001), and it plays an important role in the simulation. As specified above, mindreading lets agents address whether or not they should "trust" a particular other to act on it's "I will always cooperate" claim (cf. R. Hardin

[5] Of course, an individual *could* be better at (for example) detecting true statements than at detecting lies, and as our results will show, populations can evolve such differential abilities. Similarly, individuals might or might not be aware that they have such differential abilities, if they do. But our present *modeling* concern is with the structure of cognitive architecture, and a mechanism designed for mindreading—for detecting whether or not another individual is telling the truth—cannot be constructed on the assumption that it "knows" the answer to that question in the particular case.

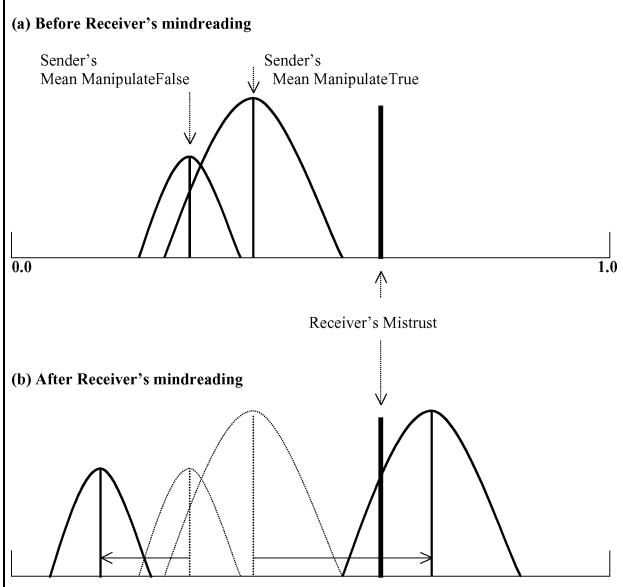
1991 and Orbell, Dawes, and Schwartz-Shea 1994). But a distinct issue—related to the concept of social capital (Putnam 2000)—concerns trust as a *generalized* disposition toward others, a willingness to accept others' "I will always cooperate" messages independent of expectations that mindreading might have produced about particular individuals, and we have incorporated this into the simulation. We model generalized "mistrust" as a value between 0.0 and 1.0, with individuals at the low end of that distribution being disposed to accept most or all of another individual's messages—whether true or false, and operationally independent of what their mindreading tells them they should accept from "this" particular individual—and individuals at the high end being disposed to *reject* most or all such messages. Only messages whose persuasiveness, modified by Sender's manipulation and Receiver's mindreading, places them above Receiver's mistrust threshold are believed by Receiver and incorporated into its decision-making.

Individual Decision-Making

An encounter between two individuals obliges each to decide between entering a PD game with the other vs. taking the alternative, non-PD course of action (which we call "ALT"). By assumption, individuals know their own cooperate vs. defect choice, and they also know the payoffs available from a joined PD game and from ALT.[6] The only additional information they need before making a decision, therefore, is the probability with which the other individual will cooperate in a joined PD game. A decision-maker's estimate of this will be a function of (1) the other individual's actual PC, (2) the other individual's capacity to manipulate both its true and its false messages, (3) the decision-maker's own mindreading capacity, and (4) the decision-maker's own level of generalized mistrust. Figure 1 uses illustrative values to show how the PD vs. ALT decision is made.

Sender is attempting to manipulate Receiver into believing its 100 "I will always cooperate" messages (although, of course, both individuals play both roles simultaneously). In this example, Sender's PC is .75, meaning that it is sending 75 true messages and 25 lies, with each set of messages distributed (normally) around Sender's manipulative capacities for true and false messages, respectively; in the example, those capacities are .40 for true messages and .30 for false ones, making our example Sender *somewhat* better at persuading with respect to truths than to lies. The best outcome for Receiver would be to reject all Sender's lies while accepting all its truths, but this receiver's mistrust threshold is set at .60, meaning that it will only accept messages that are above that threshold (absent mindreading, in the example, none). Absent mindreading, therefore, this receiver would underestimate this sender's PC, believing it to be zero rather than the true value of .75.

FIGURE 1. An Example of How Cognitive and Dispositional Attributes Interact in Receiver's Decision-Making



Note: Although this Sender's 75 true statements are somewhat more persuasive than its 25 false ones (.40, in comparison with .30), with no mindreading being exercised no message in either its true or its false sets falls above Receiver's .60 mistrust threshold. Receiver would, then, assess Sender's PC as zero. After mindreading is exercised, most—although not all—of this Sender's true statements fall above Receiver's Mistrust threshold. Those that do are used to define Receiver's estimate of Sender's PC; in this case, that estimate is somewhat less than the true .75. Of course, each of these attributes is likely to vary among individuals, with varying consequences for the accuracy of such estimates.

In general, any positive mindreading capacity on Receiver's part will lead it to read Sender's true messages as more believable than otherwise (graphically, farther to the right) and to read its lies as less believable than otherwise (farther to the left). In the Figure 1 example, Receiver's mindreading capacity of .5 leads it to read Sender's 75 true messages as having a mean believability of .7 (.5 of the distance between their manipulated persuasiveness of .4 and being fully believable at 1.0) and will lead it to read Sender's 25 lies as .15 believable (.5 of the distance between their manipulated persuasiveness of .3 and being fully unbelievable at 0.0). A perfect mindreader would have done better still, reading all Sender's true messages as 1.0 believable and all of its lies as 0.0 believable. In the example, Receiver's level of generalized mistrust is set at .60, modestly high but not

[6] In natural circumstances there is likely to be some doubt about all these values, and an adaptive problem in our ancestral past—as now—is to evaluate accurately just what *is* at stake in various games, as well as what particular game is a possibility here. Indeed, it seems likely that humans and other social animals have evolved special-purpose, domain-specific cognitive mechanisms for facilitating accurate such evaluations (Gigerenzer 1997). Given our interest in the relationship between Machiavellian intelligence and cooperative dispositions, however, we can make the simplifying assumption of perfect knowledge in these respects.

completely mistrustful. After Receiver has exercised its (also modest) mindreading capacity, its mistrust threshold is low enough that it accepts as true perhaps 60 of Sender's 75 true messages, thus estimating Sender's PC as .6. Of course, any combination of the various attributes is possible.

The bottom line, however, is the decision that Receiver makes using the information that, for better or worse, it has now gathered. Using standard Prisoner's Dilemma notation (footnote 4) and defining estPC as the number of Sender's 100 mindreading-adjusted messages falling above Receiver's own mistrust threshold, if Receiver draws cooperation from its PC it will enter a PD when

$$\text{estPC}(c) + (1 - \text{estPC})(s) > \text{ALT},$$

but otherwise will choose ALT. And should it draw defection, it will enter a PD when

$$\text{estPC}(t) + (1 - \text{estPC})(d) > \text{ALT},$$

but otherwise will choose ALT. In the particular case, of course, Receiver's choice will depend importantly on the PD payoffs and the value of ALT.

Whether or not Receiver's decision will be profitable (adaptive) depends on the quality of the information it has gathered in this encounter and on the decision that the other individual—in *its* role as receiver—makes at the same time. Notice, however, that Receiver's decision is modeled as a rational one, granted the quite probably imperfect information (Simon 1985) it has gathered about Sender's intentions. Within an evolutionary frame where payoffs are in units of fitness, this is entirely appropriate since "rationality" equates with "fitness maximization," and individuals whose choices are *not* geared in that direction will not last long. But is it inconsistent to model individuals' decisions between entering PDs and playing ALT as rational but to model the decisions of those same individuals *within* PD games as produced by "dispositions" with no basis in rationality?

It is not—granted that our concern is with the consequences of natural selection on such dispositions. As emphasized above, within an evolutionary framework, the importance of such dispositions is whether or not they contribute to an individual's adaptive success in the context of decisions about whether or not to play particular games. Rationality generally assumes that individuals have no innate dispositions, toward either cooperation or defection, basing their decisions entirely on recognizable gains and losses. But *whether* cooperative dispositions will grow and prosper or be selected against and vanish within a Machiavellian world is precisely what we are investigating here. We will return to the issue of rationality insofar as it bears on evolutionary modeling later in the paper.

Selection

In the simulation, members of a population encounter each other during a generation, with the gains or losses from those encounters functioning as "units of evolutionary fitness." All the members of a generation die after the specified number of encounters has been completed, but the more successful among them reproduce, with their offspring carrying their various cognitive and dispositional attributes into the next generation, subject to mutation. Specifically, individuals whose wealth falls below zero at the end of their generation die without reproducing, those whose wealth is positive but below twice the median reproduce once, those whose wealth is between twice the median and below three times the median reproduce twice—and so on. Carrying capacity of the ecology is limited; should the number of offspring exceed that capacity, a random lottery determines which particular agents populate the next generation.[7] Generations can involve any number of encounters, but in the simulations we report below each agent encounters each other agent twice in their particular generation, once as "Alpha," who chooses whether or not to offer a PD game, and once as "Beta," whose choice is limited to accepting vs. not accepting a PD offer, should one be made. If Alpha chooses not to offer a PD, or if Beta rejects such an offer, both make ALT.

On reproduction, mutation on all a parent's dispositional and cognitive attributes—PC, mindreading, mistrust, Manipulate True and Manipulate False—is a possibility. Because all those attributes are modeled as probabilities, mutation of fixed magnitude could easily have reached a limiting point beyond which further mutation in a particular direction would not be possible. Our solution was to define each attribute from two "constituent" integers above zero, one facilitating and one inhibiting the attribute in question. Thus, for example, PC is constructed from the integers "cooperate plus" (a positive disposition toward cooperation) and "cooperate minus" (a negative such disposition), with the individual's actual PC being defined as the proportion

$$\text{cooperate plus}/(\text{cooperate plus} + \text{cooperate minus}).$$

Similarly, "mistrust" is constructed from integers defining a disposition *not to* believe that others have cooperative intentions ("skepticism") and a disposition *to* believe that they have such intentions ("credulity"), with the attribute itself being defined as the proportion

$$\text{skepticism}/(\text{skepticism} + \text{credulity}).$$

A similar approach is taken with respect to mindreading and the two manipulative capacities. Mutation on each of the integers from which a parent's attributes are constructed can thus affect the values of attributes an offspring inherits, with potential consequences—neutral, adaptive, or maladaptive—for that offspring's success within its own generation's social environment.

[7] Carrying capacity is thus a constraint preventing population size from increasing indefinitely—as could happen if natural selection on social success were the only limiting factor. The random lottery is, thus, only conducted among individuals who are already successful in terms of natural selection. In natural circumstances, of course, immigration to unpopulated areas might have been possible for many of our early ancestors.

Mutation happens on such "constituent" attributes with the parameters "probability" and "magnitude." That is, there is some fixed probability with which any given parental attribute will mutate on being transmitted to an offspring, and some fixed magnitude (positive or negative) of those mutations that *do* happen. In the simulations to be reported, probability is set at 0.05 and magnitude at 5.0— meaning that either of the two integers from which each of an offspring's several attributes is constructed could differ from those of its parent with a probability of 0.05, and by a magnitude of plus or minus 5.0.[8] Note that although the number of offspring is influenced by a parent's relative success in its own generation, the only source of change in attributes passed from a successful parent to its offspring is mutation thus specified. The model does not, therefore, involve any parent-to-offspring learning and, certainly, no Lamarckian inheritance.

Experimental Design

In these terms, the objective is to observe how population-level values of the several individual attributes change through successive generations and the selective processes by which such changes happen. In particular, we are interested in the possibility that selection on manipulation and mindreading influences selection on cooperative dispositions—and, of course, the frequency of whatever cooperative behavior follows.

We start each simulation with a population of individuals whose cognitive and dispositional attributes militate strongly against PD games being played—specifically, PC = .10, Mistrust = .90, ManipulateTrue = .10; ManipulateFalse = .10, and Mindread = .10. At least until mutation has had a chance to change population attributes, agents will send each other "I will always cooperate" messages that are mostly lies; they will have little capacity to be persuasive (either in their many lies or in their few truths); they will have little capacity to penetrate to the truth of particular messages; and they will be so generally mistrustful that even very persuasive claims will seldom be accepted. Agents in this initial generation are, in fact, substantially *nonsocial* insofar as they lack the ability to engage in collective action, and they are also substantially *nonpolitical* insofar as their Machiavellian capacities are underdeveloped. Any cooperative equilibria that might emerge from this (perhaps "Hobbesian") world will have to be explained, but so will the process by which any initial "leap" out of this world happens.

FINDINGS

Our initial exploratory runs suggested two possibilities. (1) Selection on Machiavellian intelligence *can* both produce and sustain transitions from such a "cooperation and PD unfriendly" founding population to later populations with very high mean PC values and very frequent mutually cooperative PD games; (2) Such "cooperative transitions" happen almost exclusively under the parameters $0 \leq \text{ALT} \leq c$. Our subsequent analysis focused more systematically on exploring these possibilities and on identifying the processes underlying them.

Our approach was to run multiple simulations with PD parameters constant ($t = 15$; $c = 5$; $d = -5$; $s = -15$) but varying ALT. We first confirmed the absence—more accurately, *very* low incidence—of cooperative transitions outside the parameter range $0 \leq \text{ALT} \leq c$. In retrospect, the reasons for that pattern are clear. When the payoff from ALT is higher than from mutual cooperation, only an individual who intends defection *and* is convinced that a potential partner will cooperate will rationally enter a PD game in preference to ALT. Although one competent mindreader might make that assessment, both parties must agree before a PD is joined, and the probability of both deciding the same thing about the other, though not zero, is slim—and becomes still more so as ALT approaches t.

At the opposite extreme, when ALT ≤ 0, even a low estimate of the other's PC might return an expected value for entering a PD that is higher than ALT, and because cooperation is dominated by defection, such an estimate will be particularly likely for intending defectors—probabilistically, individuals with low PC values. Estimating the other's PC at 0.4, for example, an intending defector would calculate the EV of a PD as 3, whereas an intending cooperator would calculate it at −7. Trapped between two alternatives both offering a loss, the intending cooperator—probabilistically, an individual with a *high* PC value—cannot prosper, meaning negative selection on high PC. But a population of low PC agents produces, for the most part, either PDs played to mutual defection (a negative) or ALT (also a negative in this range) and will rapidly die out.[9]

With this understanding, we ran 90 simulations, each consisting of 20,000 generations, each with the same PD payoffs but including 10 at each 0.5 interval across the range $0 < \text{ALT} < 5$. The results are reported in Table 1. First, the data confirm that transitions from overwhelmingly ALT choices to overwhelmingly PD games played with mutual cooperation *are* very frequent within that parameter range. Although there are slightly fewer transitions toward the lower end of that range, across that whole range transitions *fail* to happen in only 9 of the 90 simulation runs. Within each of the nine ALT categories, there is considerable variation as to when cooperative transitions begin, but such transitions normally do take several thousand generations to get under way (column 3). Nevertheless, and despite occasional lapses that are reversed with the passage of

[8] Given that there are 10 components on which mutation might happen, these parameters mean that there is a .599 probability of any given offspring having no mutation on any of the components it inherits from its parent. Sensitivity testing shows that increasing or decreasing these parameters does have predictable effects on the *speed* with which mean population attributes change but not on the basic pattern of findings we report.

[9] With different PD parameters the population might not die out; if mutual defection were a positive, for example, a population of very low PC types could survive indefinitely even with a negative ALT. The important point is that, with a negative ALT, cooperative dispositions will be strongly selected against, and cooperative behavior will be extremely rare—at best.

TABLE 1. Cooperative Transitions within the Parameter Range 0 < ALT < *c*

ALT Value (1)	Number of Cooperative Transitions (2)	Mean Generation Where Transition Starts (3)	Mean PC Value After the Transition (4)	Predicted Threshold PC (5)
4.5	9/10	6535	0.981	0.975
4.0	10/10	5923	0.965	0.950
3.5	10/10	5983	0.940	0.925
3.0	9/10	5924	0.926	0.900
2.5	9/10	6046	0.898	0.875
2.0	10/10	4474	0.897	0.850
1.5	8/10	3398	0.858	0.825
1.0	8/10	4040	0.833	0.800
0.5	8/10	5566	0.815	0.775

Note: Based on 10 simulations for each value of ALT.

time, the 81 cooperative transitions all produced quite high mean PC values (column 4) that persist in apparent equilibrium until the simulations end at the twenty thousandth generation.[10]

Further, those high PC values are reflected in actual behavior. We illustrate this with an example when ALT was set at 4. Figures 2a, 2b, and 2c show, respectively, mean levels of PC across the 20,000 generations of this simulation; the incidence of ALT outcomes vs. joined PD games; and, among the PD games that were joined, the incidence of mutual defection, mutual cooperation, and the outcome in which one defecting individual exploits the other. From Figure 2a, the transition to high levels of PC begins at generation 5,319 after what appear to be a couple of "false starts" but—after a thousand or so generations of some fluctuation—has settled down to a basically stable state around PC = .97 with only minor fluctuations from generation to generation. After the transition, in other words, this population is *very* disposed to cooperation, granted not *perfectly* so.

From Figure 2b, we see that these cooperative dispositions are paralleled by a willingness to enter PD games and a corresponding decline in the incidence of ALT outcomes.[11] There is some instability from generation to generation and one occasion (around generation 15,000) when the preponderance of PD games in the ecology is temporarily challenged by ALT, but the reversal in the overall pattern of game choices after the transition is clear.

Finally, from Figure 2c we see that the PD games that dominate relationships among individuals after the transition have, overwhelmingly, mutually cooperative outcomes. The one fluctuation that does occur is associated with the *incidence* of PD games, not with any major fluctuation of within-game behavior, and mutual cooperation remains the characteristic outcome to PD games throughout this period. Importantly, notice that the "trickle" of PD games that we observe after about generation 1,000 (Figure 2b) is identified here as being games played almost exclusively to mutual defection. In fact, only two mutually cooperative games were played during the 5,318 generations before the cooperative transition got under way. We will return to the significance of this "trickle" of mutually defecting PD games shortly. In general, however, these data show that the rapid change in cooperative dispositions is associated with a corresponding change in behavior from overwhelmingly ALT choices to overwhelmingly joined PD games and, within those games, frequent cooperation. This granted, the problem is to explain (a) the origin of such cooperative transitions and (b) the processes by which they are maintained at what appears to be equilibrium.

Origins

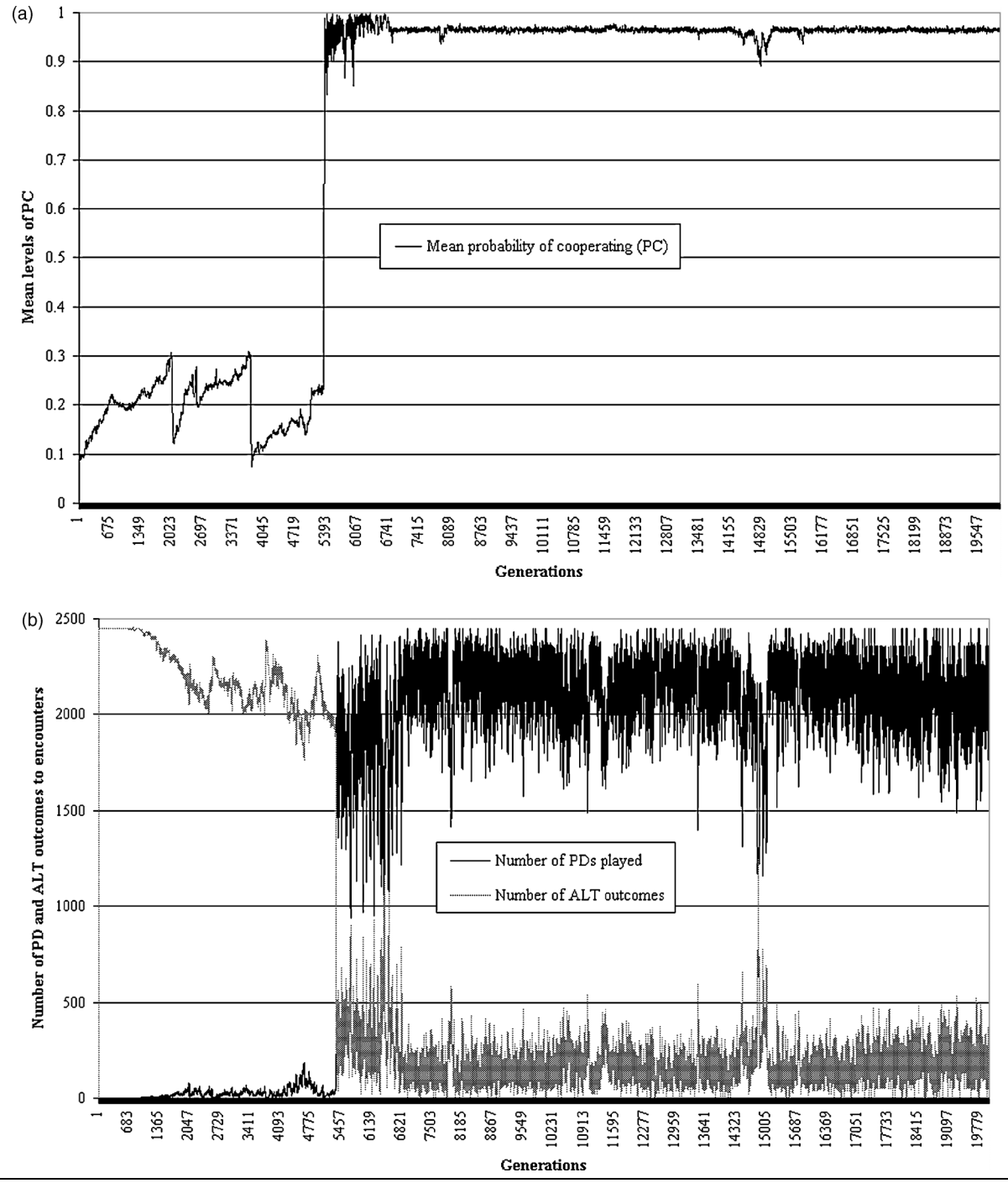
To find answers, we examine mean levels of the three Machiavellian attributes (ManipulateTrue, ManipulateFalse, Mindread), the generalized disposition to mistrust, and cooperative dispositions, as those varied with PD games played. Although population changes are the ultimate concern of evolutionary analyses, aggregate data do not speak directly to the selective pressures on individuals. Therefore, we will also report correlational data at the individual level during adaptively important periods.

Of greatest initial interest, Figure 3 shows, for the case being followed, a substantial step-level, upward change in mean population mindreading values at around generation 2,000, a change that is plausibly a response to the costs of becoming involved with PD games when the mean PC is very low, thus the probability of taking a negative from such a game is high—whether as a (rare) cooperator or as a (more frequent) defector. A strong test of that hypothesis is provided by observing the relationship between individuals' mindreading scores (independent variable) and the proportion of PD games that, if offered to an individual, were accepted, across the 100 generations between 1,990 and 2,090 when the increase in mindreading was most marked and the variance in mindreading was also

[10] More extensive testing showed that the general pattern of cooperative transitions happening only within the specified parameter range was robust—when, for example, the initial world was cooperation-*friendly* and across wide variation in the absolute values of PD payoffs.

[11] The lack of a perfect negative match between PD and ALT outcomes is explained by there being two ways the latter outcome can happen—when Alpha (the first mover) rejects the PD option and when Beta (the second mover) rejects a PD offer from Alpha. PD games, on the other hand, require both parties to agree.

FIGURE 2. A Cooperative Transition When ALT = 4—Mean PC and Behavioral Outcomes: (a) Mean Levels of PC, (b) ALT vs. PD Outcomes to Encounters, and (c) Outcomes to Joined PD games

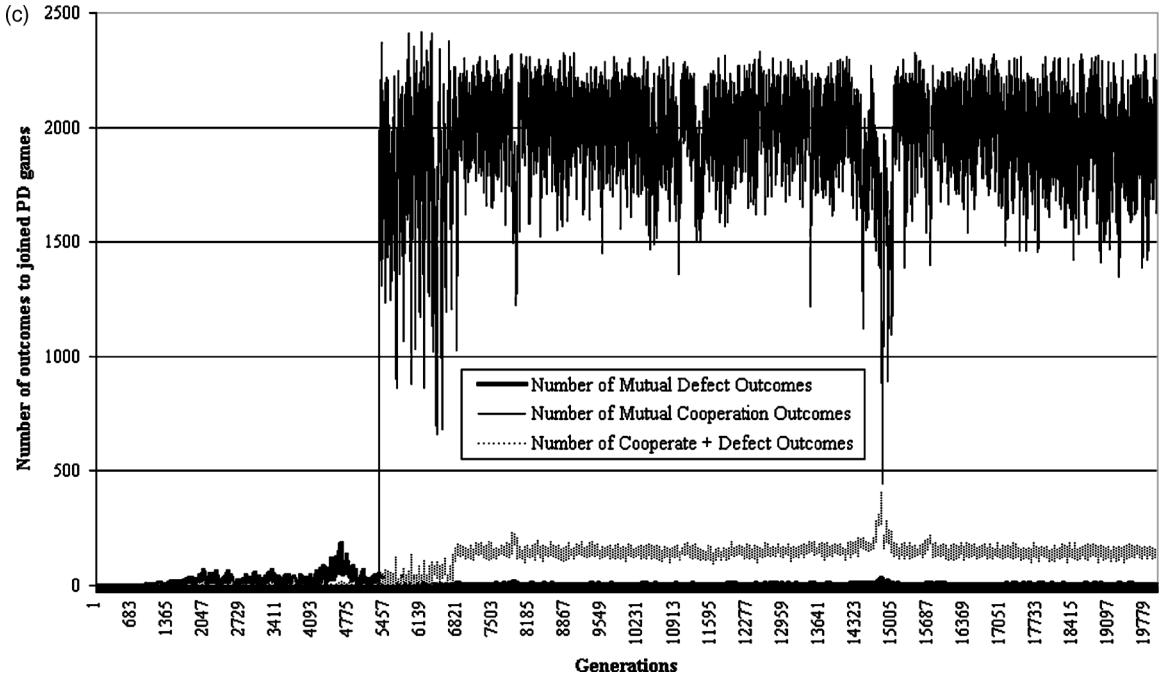


highest.[12] The result was $\beta = -0.1593$ ($p < .001$), supporting the hypothesis that selective pressures in this nasty period were favoring individuals with mindreading capacities sufficient to keep them from accepting invitations to enter PD games—that, *if* entered, would have resulted in loss.

Notice that positive selection on PC values in this early, cooperation-unfriendly stage is not likely because so few PD games are being played—with those that are played resulting in costly mutual defection. But what happens to individuals whose PC does happen to mutate upward? We can predict, first, that such "cooperative mutants" will be *spotted* as potential victims by mindreading-equipped low-PC types, and the data support this: In the case being followed, for example,

[12] Data points are each of the 50 individuals in each of the 100 generations, for an n of 5,000.

FIGURE 2—*Continued*.



the relationship between agents' own PC values (independent variable) and the number of PD games offered to them by others for the 100 generations immediately prior to the start of the cooperative transition was $\beta = 36.3824$ ($p < .0001$). The same mindreading capacity that evolved to address the prophylactic function of keeping agents out of costly PD games in a cooperation-unfriendly world also equips them to recognize and try to exploit occasional high PC types. But mindreading cuts two ways, and Table 2 reports

FIGURE 3. A Cooperative Transition When ALT = 4—Mindreading

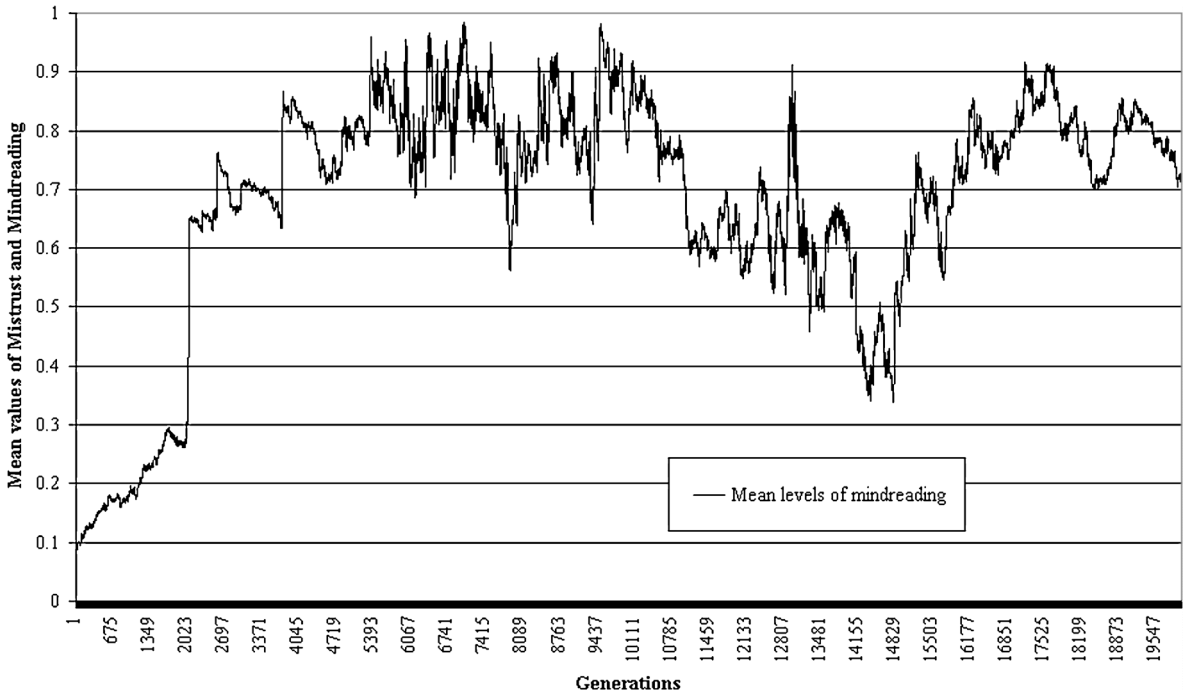


TABLE 2. Multiple Regression Beta Weights: Proportion of Offered PD Games Accepted by Mistrust, Mindreading, and PC

	Coefficient
Intercept	1.2216
Mistrust	−0.6821*
Mindreading	−0.9414*
Probability of cooperating	−0.0641*

Note: One hundred generations prior to the cooperative transition. $R^2 = .557$; $^*P < .001$; $N = 4{,}978$. Observations are on 50 agents in each of the 100 generations; missing cases had zero offers.

coefficients from a multiple regression predicting, from the PC, mindreading, and mistrust scores of such target agents (independent variables), the proportion of PD offers that, once made, were *accepted* by agents. Clearly, the mindreading capacities of such occasional "cooperative mutants" makes it possible for them to recognize and avoid such traps—thus, for their relatively higher cooperative dispositions to be passed on to any offspring they might have.

Granted, sustained upward drift of PC values will require low-probability upward mutations on the PC values of at least several members of a lineage, something that is not likely for any particular lineage. But mindreading does make such "cooperative drift" at least *possible* and, with the passage of many generations, quite likely on at least *some* lineages. And this is what sets the stage for the observed cooperative transitions.

We have followed the individual-by-individual details across several different values of ALT and—with due allowance for the variation resulting from the many stochastic elements in the simulation—the basic pattern is always the same. At some point, two relatively cooperative agents, both equipped with high mindreading, each recognize the other as a *good* bet for a PD game, and with one offering and the other accepting, both capture the mutual cooperate payoff. With other agents all rejecting PD games in favor of ALT (where ALT $< c$), these two enjoy relative reproductive success. Their offspring, inheriting the high PC and high mindreading that made their parents successful, have similar success in their own generation until, quite rapidly, the whole population is descended from this original pair.

Because this process relies on randomly drifting PC values, such cooperative transitions are not inevitable within this parameter range. Quite possibly, no two individuals will ever reach high enough PC levels to be seen as attractive partners by each other. In fact, as we have pointed out, transitions failed in nine of the 90 simulation runs we conducted. By the same token, some transitions might take many more generations to happen than others. But the process we have described makes cooperative transitions *likely* to happen—sooner or later.

Returning to Table 1, it is apparent that the level at which PC equilibrates varies monotonically with ALT; as ALT becomes higher within the critical parameter range, mean PC values after the transition also become higher. Because natural selection can be assumed to "discover" attribute levels that are optimal for particular environments, it appears that optimal levels of cooperative dispositions are positively related to ALT. This is perhaps surprising, as one might expect cooperative dispositions to be *higher* as the relative advantage of mutual cooperation over other ways of making a living becomes greater, but just the opposite happens.

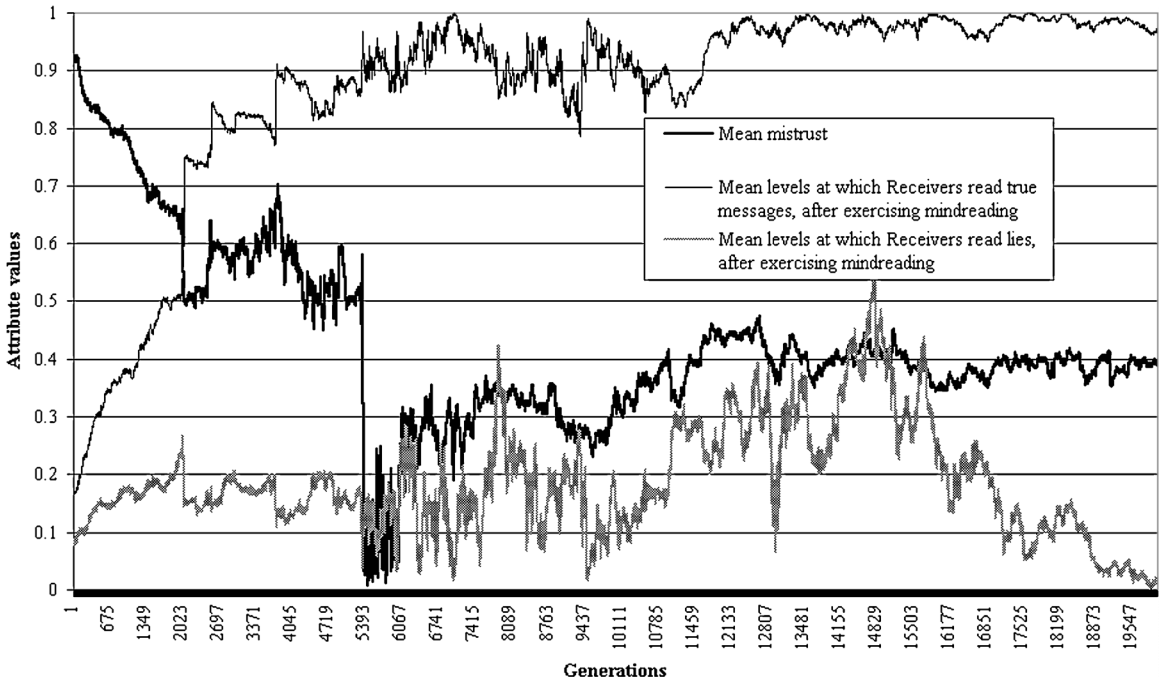
The issue is resolved by recognizing that two criteria must be met in order to maximize returns in this world. First, an agent must persuade another to *enter* a PD game rather than to play ALT; and, second, under that constraint, it must maximize its own returns *within* any PD game that might be joined. Given the potential victim's (now) probably high mindreading capacities, meeting the first criterion requires the agent to have *genuinely* high cooperative dispositions; when ALT is high, a potential victim can afford to be fussy about the cooperative dispositions of would-be victimizers. But the would-be victimizer's cooperative dispositions should be no higher than minimally necessary to get the potential victim into play. Being any *more* cooperatively disposed will result in lost opportunities for profitable free riding, placing the agent at an evolutionary disadvantage relative to others who do manage to enter PD games with cooperative individuals but who *do*, nevertheless, defect.[13]

This analysis suggests: *Natural selection will "discover" levels of cooperative dispositions that are simultaneously high enough to ensure that prospective partners assess the expected value of entering PD games as greater than ALT, but low enough to maximize the possibility of exploiting partners should a PD game be joined*. For a population of perfect mindreaders, this logic predicts a threshold PC for each value of ALT (just) above which a population's PC values will equilibrate—specifically where $\mathrm{PC}(c) + (1 - \mathrm{PC})(s) = \mathrm{ALT}$. In these terms, column 5 in Table 1 specifies such thresholds for the nine values of ALT that we have examined across the parameter range $0 < \mathrm{ALT} < c$, and Column 4 shows that, as predicted, the mean levels at which cooperative dispositions equilibrate after transitions are (1) monotonically related to values of ALT within this range and (2) only slightly above the threshold in each case—the slight "error" in each case being, presumably, a consequence of minor mistakes in mindreading that still do happen.[14]

[13] Of course, in this situation two individuals encountering each other are *both* would-be victimizers and potential victims.

[14] Although transitions do frequently *happen* across all values of ALT within this parameter range, case-by-case and generation-by-generation micro analyses of the *process* by which they happen show that there are many more "false starts" at the lower values of ALT where several lineages normally appear about to dominate the entire population but die out before doing so. Whereas one normally does, sooner or later, do so, the reason for such false starts is the permissiveness that low values of ALT provide for relatively defection-inclined individuals who are more likely to attempt exploitation of each other than their (necessarily) more cooperative counterparts when ALT is high. Correspondingly, the permissiveness provided for defection at very low values of ALT makes cooperative transitions, when they do happen, more unstable than at higher such values, in a very small number of cases allowing them to collapse altogether.

FIGURE 4. A Cooperative Transition When ALT = 4; Mean Mistrust and Mean "Mindreading-Adjusted" Truths and Lies



Maintenance

What, then, sustains such cooperative transitions once they have happened? From Figure 4, we observe a rapid drop in mistrust as the cooperative transition is getting under way (generation 5,319). This is explained in terms of competition *among* individuals in the newly cooperative population. In this "nice" world, advantage will go to those who enter more PD games than their adaptive competitors, and with mindreading already high, this means selective advantage to those who have relatively low levels of mistrust. The result is that, for a few thousand generations, the posttransition population is not only highly disposed to cooperate, but also very low in mistrust.

But such trusting dispositions in combination with high PC values make this population *doubly* ripe for exploitation by low PC types, thus for an eventual reversion to generally low PC values. The reason why such a reversion does not happen involves the relationship between mean levels of mistrust and mindreading. First, from Figure 4 we can see that the drop to very low levels of mistrust only lasts a few thousand generations after the transition has happened. Although that drop did facilitate individuals' entering now generally cooperative relationships, it also opened the door for relatively *less* cooperative individuals to prosper and spread, with that, in turn, selecting for individuals with *somewhat* higher mistrust, as reflected in the mean values in Figure 4. The adaptive problem is to find an *optimal* level of mistrust, given widespread mindreading capacities but also the substantial gains that are now available from entering PD games in this generally cooperative world.

The solution "discovered" by natural selection is a balance between mistrust and mindreading, with mistrust *high* enough to ensure that lies are rejected but *low* enough not to endanger acceptance of the true messages that predominate in this posttransition, generally cooperative world. Figure 4 illustrates how this happens using the case we have been following. Granted the elevated mindreading that is now normal, receivers accept senders' true messages with a higher probability than the persuasiveness of those messages would imply, but they also accept Senders' lies with a *lower* probability than the persuasiveness of *those* messages would imply. Critically, mean levels of mistrust evolve upward to a point that is high enough to ensure that Senders' lies are rejected, thus precluding invasion by lower PC types. There are two brief periods where the mean "mindreading-adjusted persuasiveness" of lies moves above mean mistrust, and the consequences of that are visible in the brief drop in the incidence of cooperatively played PD games (Figures 2b and 2c). But in both cases this drop is only temporary and the cooperative equilibrium returns.

The reason for this equilibrium is that the most adaptive configuration of cognitive and dispositional attributes is one that will maintain frequent access to PD games played with cooperators. But in general, advantage will go to those whose mistrust is high enough, in conjunction with their mindreading capacity, to ensure rejection of most, if not all, lies that are directed toward them, while being low enough to ensure acceptance of

most, if not all, the truths that are also directed toward them. Those who best strike such an optimal balance will be best positioned to prosper within a population of individuals who are, in general, strongly disposed to cooperative behavior but who nevertheless *sometimes* do defect.

In summary, in the initial cooperation-unfriendly world, individuals can occasionally be led to enter PD games that are played to costly mutual defect outcomes, thus selecting for high mindreading capacities. Once in place, those capacities permit upward drift on cooperative dispositions. Individuals with such dispositions are the frequent targets of exploitative attempts by those who are more defection-inclined, but their mindreading allows them to survive (via ALT choices) no differently from others in the population. At some point, however, two such high PC individuals are likely to play a PD game with each other and to do so cooperatively—and, given that ALT $< c$, they will prosper by comparison with the rest of the population. Quite rapidly, the descendants of these individuals populate the entire ecology. The equilibrium level of cooperative dispositions after such a transition is high enough to attract skilled mindreaders into PD games but low enough to extract some gains from exploiting them once they have entered. Transitions are sustained by selection in favor of individuals who "discover" an optimal mixture of mindreading and mistrust—sufficient to ensure that they accept the true "I will always cooperate" messages characteristic of this environment but low enough to ensure that they reject the false messages that are, nevertheless, still being sent.

DISCUSSION

As is often pointed out, humans' ancestors were almost certainly cooperative animals well before the point at which our line diverged from that of other large primates (e.g., Caporael et al. 1989), meaning that our "cooperation-unfriendly" starting world is best understood as a convenient analytic device for showing how cooperative dispositions could evolve despite adaptive pressures to the contrary. And if cooperative transitions resembling those we have discussed ever *did* happen, there can be no implication that they happened with anything approximating the speed across natural generations that we observe in the "generations" of our simulation. Any natural cooperative transitions could well have taken hundreds of thousands of years.

All we claim to have shown is that dispositions to cooperate *can* evolve and be sustained at equilibrium as a direct consequence of selection on Machiavellian capacities for manipulating the content of messages sent to others and for mindreading to the truth underlying others' attempts at manipulation. As sketched above, the literature on social evolution has identified a number of plausible evolutionary paths to cooperative behavior, the best known being kin selection (Hamilton 1964), reciprocity (Trivers 1971), and group selection (Sober and Wilson 1998). And the broad idea that humans and other highly social animals have evolved "Machiavellian" capacities supporting their capacity to successfully negotiate the competitive challenges of group life is a widespread and peculiarly fertile one. To our knowledge, however, there has been no model of how our cooperative dispositions and our Machiavellian capacities for manipulation and mindreading might be functionally related, and that is the gap we have attempted to fill here.

Although the problem of cooperation has a long analytic history in political science, the idea of Machiavellian intelligence—that we are a "political animal" in our cognitive design as well as in the fact that we relate to each other in groups—is less well known, and we believe that it deserves greater currency. This is certainly the case if, as our data suggest, Machiavellian intelligence and cooperative dispositions have an intertwined evolutionary history.

Our model is based on natural selection acting on individual cognitive and dispositional attributes, but the processes it identifies could interact with cultural evolution as modeled, for example, by Boyd and Richerson (1985). As we have shown, cooperative transitions occur when a pair of agents with high cooperative dispositions and well-developed mindreading capacities recognize each other as offering a good PD bet and, joining such a game, both cooperate and prosper accordingly. In our pared-down world, however, agents do not have the capacity to learn from others' experience, to adapt their behavior by observing others' success, but such a capacity certainly does exist among humans and, indeed, among some other primates (see, e.g., de Waal 2001). Incorporating that capacity, we might see the success of cooperative behavior among others being recognized by individuals not personally disposed in that direction, with the behavior spreading by imitation—independent of the cognitive and dispositional evolution that concerns us here.

This raises many further possibilities. For example, populations could evolve to include some individuals whose cooperation is a consequence of genetically based cooperative dispositions and others whose cooperation is a consequence of social learning coupled with a "strategic" recognition that cooperative behavior will work for them as it has for others. More interesting, perhaps, we might see individuals appearing whose cooperative behavior is a product of both innate dispositions *and* such social learning. Social learning could also reveal opportunities for Machiavellian exploitation that would not be recognized in its absence, further accelerating the evolutionary arms race between manipulation and mindreading. And benefits received as a consequence of social learning could select for the (presumably genetically based) *capacity* for social learning—perhaps in an upward spiral involving that capacity, cooperative dispositions, and Machiavellian manipulation and mindreading. But these are speculations, taking us well beyond our more modest present concern with the evolutionary relationship among just the latter two attributes.

The idea of rationality is presently under dispute in the discipline (Green and Shapiro 1994), but the evolutionary model proposed here allows us to distinguish between two versions of rationality, perhaps

contributing to a reconciliation between the idea of rationality and the empirical data on cooperation—mentioned above—that has featured in that dispute.

Rationality in action: This is the standard way in which rationality is employed in political science and related disciplines, notably economics. Essentially, individuals choose so as to maximize their private welfare—under a variety of constraints, most importantly, on information. Within our simulation, this is how we have modeled individual actors choosing between entering PD games and playing ALT. Their choices are constrained by others' success at manipulation and by their own limits with respect to mindreading, but they do the best they can. And, we believe, that is how individuals *should* be modeled in an evolutionary simulation such as ours. As Alchian (1950) long ago pointed out, the idea that individuals choose *as if* they were rational in this classic sense can be employed as a profitable fiction as long as interest resides in equilibria that are produced through an evolutionary process that selects for fitness maximizing behavior. Remembering that payoffs in our evolutionary simulation are "units of fitness," individuals who do *not* act in such a manner would rapidly be selected out, making it reasonable to start from this model of action, even if it were in no way parallel to the actual deliberative or cognitive processes that ancestral populations employed in their social decision making.

Yet, as we have pointed out, this is *not* how we have modeled choices by those same individuals within any PD games that might be joined. Those are modeled as random draws from a probability distribution in order to capture the idea of cooperative *dispositions* that have no necessary basis in the individual's calculation of self-interest or, indeed, in its actual self-interest. By this emphasis on dispositions unrelated to calculations of current interest, we are departing from the tradition in political science and elsewhere that seeks solutions to the "problem of cooperation" in processes that are founded in "rationality in action."

Rationality in Design: Here rationality refers to the adaptive fit between some designed apparatus and the environmental problems that apparatus is intended to solve—an idea usefully captured within an evolutionary context by Tooby and Cosmides' (1992) metaphor of an appropriately designed key being one that opens a particular lock. As these authors propose, a focus on design requires asking a series of *engineering* questions, most importantly about the correspondence between the problem to be solved and the mechanism "designed" by natural selection to solve it. Granted that a particular adaptive problem was a repetitive part of the ancestral environment, what design solutions has natural selection produced in response? And thinking as a "reverse engineer" (Dennett 1995), What adaptive pressures in the ancient environment are most likely to have led natural selection to "design" particular complex structures—presumed "adaptations"—that we observe today?

This "rationality in design" approach lets us address the empirical fact of frequent cooperative behavior that is anomalous in terms of rationality in action. *Our simulation suggests that the most adaptive configuration of cognitive and dispositional attributes—the most rational design response by natural selection to the problems of group living—is strongly but not perfectly cooperative dispositions, a modest but not particularly high level of mistrust, and a substantial ability to mindread as well as to manipulate.* Such a model not only is consistent with the laboratory data about cooperation, but also has the advantage of explaining cooperative dispositions squarely within the broader intellectual (and interdisciplinary) enterprise seeking an evolutionary understanding of humans' cognitive adaptations for social life, including Machiavellian capacities.

This model is not inconsistent with *some* individuals having only very weak dispositions to cooperate, just as it is not inconsistent with some having unusually strong, even perfect, such dispositions. Individual differences aside, however, the model addresses *species-typical* attributes, suggesting that, as a default, we should expect social animals—most interestingly, of course, humans—to be quite strongly disposed to cooperative behavior, as well as modestly trusting and reasonably adept at both manipulation and mindreading. How individuals' cooperative dispositions might be undermined or reinforced by the incentive attributes of the present situation is, of course, a different matter.

The availability of an alternative to playing prisoner's dilemma games as a way of gathering resources is critical to our model, and our finding that cooperative transitions only happen within the finite range $0 < \text{ALT} < c$ invites further study. Whereas we believe that the logic by which transitions are so confined is clear, the more difficult question concerns the natural world circumstances that might embody that logic—or, more accurately perhaps, might *have* embodied that logic in the ancient environment. Most interesting, we think, is the finding that cooperative dispositions evolve to their highest level when the payoff from alternative ways of gathering resources most closely approaches the payoff from mutual cooperation without actually exceeding it. This appears counter-intuitive at first; one might expect cooperation to evolve to its highest levels when there is the *most* to be gained from mutual cooperation relative to alternative courses of action. But as the argument about particular equilibria of cooperative dispositions suggests, a greater distance between ALT and c just allows more "room" for *non*cooperative dispositions to evolve.

We recognize, of course, that the ancestral environment that shaped our modern cognitive apparatus did not involve a single set of payoff parameters (as do particular runs of our simulation), but a substantial diversity of particular parameters, making that environment best thought of as a statistical average across that diversity (Daly and Wilson 1999; Tooby and Cosmides 1990). Nevertheless, our model does provide a basis for hypothesizing that sociality itself—a willingness to enter PD-type games coupled with a strong disposition to play such games in a cooperative manner—evolved to its highest levels when there were only marginal gains to be had from jointly cooperative action in comparison

with "going it alone." The problems associated with sociality became more acute as relatively higher payoffs became available from cooperative activities, thus permitting the evolution of at least somewhat more ambivalent cooperative dispositions.

REFERENCES

Alchian, Armen. 1950. "Uncertainty, Evolution and Economic Theory." *Journal of Political Economy* 58 (3 June): 211–21.

Axelrod, Robert. 1981. "Emergence of Cooperation among Egoists." *American Political Science Review* 75 (June): 306–18.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Boehm, Christopher. 1997. "Egalitarian Behavior and the Evolution of Political Intelligence." In *Machiavellian Intelligence II*, ed. R. W. Byrne and A. Whiten. Cambridge: Cambridge University Press, 341–64.

Bowlby, J. 1969. *Attachment and Loss*. New York: Basic Books.

Boyd, Robert, and Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Sciences of the United States of America* 1000 (March 11): 3531.

Brothers, Leslie. 1997. *Friday's Footprint: How Society Shapes the Human Mind*. New York: Oxford University Press.

Byrne, Richard W., and Andrew Whiten, eds. 1988. *Machiavellian Intelligence; Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. New York: New York: Oxford University Press.

Camerer, C. F., and R. Thaler. 1995. "Anomalies—Ultimatums, Dictators, and Manners." *Journal of Economic Perspectives* 9 (Spring): 209–19.

Caporael, Linnda, Robyn Dawes, John Orbell, and Alphons van de Kragt. 1989. "Selfishness Examined: Cooperation in the Absence of Egoistic Incentives." *Behavioral and Brain Science* 12 (December): 683–99.

Daly, Martin, and Margo Wilson. 1999. "Human Evolutionary Psychology and Animal Behavior." *Animal Behaviour* 57 (March): 509–19.

Dawes, Robyn. 1975. "Formal Models of Dilemmas in Social Decision-Making." In *Human Judgment and Decision Processes*, ed. M. F. Kaplan and S. Schwartz. New York: Academic Press, 87–108.

Dawkins, Richard. 1976. *The Selfish Gene*. New York: Oxford University Press.

Dawkins, Richard, and John R. Krebs. 1978. "Animal Signals: Information or Manipulation?" In *Behavioural Ecology: An Evolutionary Approach*, ed. J. R. Krebs and N. B. Davies. Oxford: Blackwell Scientific, 282–309.

de Waal, Frans. 2001. *The Ape and the Sushi Master; Cultural Reflections by a Primatologist*. New York: Basic Books.

Dennett, Daniel. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.

Dobzhansky, Theodosius. 1973. "Nothing in Biology Makes Sense Except in the Light of Evolution." *American Biology Teacher* 35 (March): 125–29.

Dunbar, Robin I. M., Chris Knight, and Camilla Power, eds. 1999. *The Evolution of Culture*. New Brunswick, NJ: Rutgers University Press.

Field, Alexander J. 2001. *Altruistically Inclined? The Behavioral Sciences, Evolutionary Theory, and the Origins of Reciprocity*. Edited by T. Kuran, *Economics, Cognition, and Society*. Ann Arbor: University of Michigan Press.

Frohlich, N., and J. Oppenheimer. 1996. "Experiencing Impartiality to Invoke Fairness in the N-Pd: Some Experimental Results." *Public Choice* 86 (January): 117–35.

Fukuyama, Francis. 1995. *Trust: The Social Virtues and the Creation of Prosperity*. New York: Free Press.

Gigerenzer, Gerd. 1997. "The Modularity of Social Intelligence." In *Machiavellian Intelligence II: Extensions and Evaluations,* ed. A. Whiten and R. W. Byrne. Cambridge: Cambridge University Press, 264–88.

Green, Donald P., and Ian Shapiro. 1994. *Pathologies of Rational Choice Theory; A Critique of Applications in Political Science*. New Haven, CT: Yale University Press.

Hamilton, W. D. 1964. "The Genetical Evolution of Social Behaviour. I and II." *Journal of Theoretical Biology* 7 (July): 1–52.

Hardin, Garrett. 1977. *The Limits of Altruism; An Ecologist's View of Survival*. Bloomington: Indiana University Press.

Hardin, Russell. 1991. "Trusting Persons, Trusting Institutions." In *The Strategy of Choice*, ed. R. J. Zeckhauser. Cambridge, MA: MIT Press, 185–209.

Hobbes, Thomas. [1651] 1947. *Leviathan*. New York: E. P. Dutton.

Humphrey, Nicholas K. 1976. "The Social Function of Intellect." In *Growing Points in Ethology*, ed. P. P. G. Bateson and R. A. Hinde. Cambridge: Cambridge University Press, 303–17.

Jolly, Allison. 1966. "Lemur Social Behavior and Primate Intelligence." *Science* 153 (July): 501–6.

Mithen, Steven. 1996. *The Prehistory of the Mind; The Cognitive Origins of Art, Religion and Science*. London: Thames & Hudson.

Nesse, Randolph M. 2001. *Evolution and the Capacity for Commitment*. ed. K. S. Cook, M. Levi and R. Hardin, *The Russell Sage Foundation Series on Trust*. New York: Russell Sage Foundation.

Olson, Mancur. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Orbell, John, and Robyn Dawes. 1993. "Social Welfare, Cooperators' Advantage and the Option of Not Playing the Game." *American Sociological Review* 58 (December): 787–800.

Orbell, John, Robyn Dawes, and Peregrine Schwartz-Shea. 1994. "Trust, Social Categories and Individuals: The Case of Gender." *Motivation and Emotion* 18 (June):109–28.

Orbell, John, Robyn Dawes, Randy Simmons, and Alphons van de Kragt. 1986. "Organizing Groups for Collective Action." *American Political Science Review* 80 (December): 1171–85.

Orbell, John, Tomonori Morikawa, and Nicholas Allen. 2002. "The Evolution of Political Intelligence: Simulation Results." *British Journal of Political Science* 32 (October): 613–39.

Ostrom, Elinor. 1998. "A Behavioral Approach to the Rational Choice Theory of Collective Action." *American Political Science Review* 92 (March): 1–22.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without the Sword: Self-Governance Is Possible." *American Political Science Review* 86 (June): 404–17.

Putnam, Robert. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.

Rapoport, Anatol, Melvin J. Guyer, and David G. Gordon. 1976. *The 2×2 Game.* Ann Arbor: University of Michigan Press.

Reeve, Hudson Kern. 2000. "Multi-Level Selection and Human Cooperation." *Evolution and Human Behavior* 21 (January): 65–72.

Simon, Herbert. 1985. "Human Nature in Politics; The Dialogue of Political Science with Psychology." *American Political Science Review* June (2): 293–304.

Sober, Elliott, and David Sloan Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Tinbergen, N. 1963. "On Aims and Methods of Ethology." *Zeitschrift fur Angewandte Zoologie* 20: 410–33.

Tooby, John, and Leda Cosmides. 1990. "The Past Explains the Present: Emotional Adaptations and the Structure of Ancestral Environments." *Ethology and Sociobiology* 11 (4–5): 375–424.

Tooby, John, and Leda Cosmides. 1992. "The Psychological Foundations of Culture." In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, ed. J. Barkow, L. Cosmides, and J. Tooby. New York: Oxford University Press, 19–136.

Trivers, Robert. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46 (March): 35–57.

Trivers, Robert. 1985. *Social Evolution*. Menlo Park, CA: Benjamin/Cummings.

Whiten, Andrew, and Richard W. Byrne, eds. 1997. *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge: Cambridge University Press.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, NJ: Princeton University Press.

Wynn, Thomas. 2002. "Archaeology and Cognitive Evolution." *Behavioral and Brain Science* 25 (June): 389–438.