KNOWING THYSELF: THE EVOLUTIONARY PSYCHOLOGY OF MORAL REASONING AND MORAL SENTIMENTS

Leda Cosmides and John Tooby

A*bstract*: "Ought" cannot be derived from "is," so why should facts about human nature be of interest to business ethicists? In this article, we discuss why the nature of human nature is relevant to anyone wishing to create a more just and humane workplace and society. We begin by presenting evolutionary psychology as a research framework, and then present three examples of research that illuminate various evolved cognitive programs. The first involves the cognitive foundations of trade, including a neurocognitive mechanism specialized for a form of moral reasoning: cheater detection. The second involves the moral sentiments triggered by participating in collective actions, which are relevant to organizational behavior. The third involves the evolved programs whereby our minds socially construct groups, and how these can be harnessed to reduce racism and foster true diversity in the workplace. In each case, we discuss how what has been learned about these evolved programs might inform the study and practice of business ethics.

Introduction: Human Nature *and* Ethics

Human beings have moral intuitions. Assume, for a moment, that some of these reflect the operation of reliably developing neural circuits, which implement programs that are species-typical and therefore cross-culturally universal. That is, assume that some forms of moral reasoning and moral sentiment are produced by elements of a universal human nature. Does this justify them ethically?

Of course not. Natural selection favors designs on the basis of how well they promote their own reproduction, not on how well they promote ethical behavior. If this is not obvious, consider the fate of a mutation that alters the development of a neural circuit, changing its design away from the species standard. This new circuit design implements a decision rule that produces a radically different ethical choice in a particular type of situation: help rather than hurt, cooperate rather than free ride. Will this new decision rule, initially present in one or a few individuals, be eliminated from the population? Or will

it be retained, increasing in frequency over the generations until it replaces the old design, eventually becoming the new species standard?

The fate of the mutant decision rule will be jointly determined by two ethically blind processes: chance and natural selection. Chance is blind not only to ethics, but to design: it cannot retain or eliminate circuit designs based on their consequences. Natural selection, however, is not blind to design. The mutant design and the standard design produce different ethical choices; these choices produce different consequences for the choosers, which can enhance or reduce the average rate at which they produce offspring (who carry the same design). If the mutant decision rule better promotes its own reproduction (through promoting the reproduction of its bearers), it will be favored by selection. Eventually, over the generations, it will become the new species-standard. The decisions it produces—ethical or otherwise—will become the "common sense" of that species.

This is the process that, over eons, constructed our human nature. As a result, human nature is comprised of programs that were selected for merely because they outreproduced alternative programs in the past. There is nothing in this process to ensure the production of decision rules or moral sentiments that track the desiderata of an ethically justifiable moral system. So why should ethicists care about human nature?

Human nature matters for three reasons. First, outcomes matter. Many ethicists are concerned with how to create a more just and humane society, starting in the workplace. But what policies are capable of achieving this? Whereas some moral philosophers argue that an outcome is ethical if the procedure that produced it was ethical (e.g., Nozick, 1975), others argue that certain outcomes are ethically better than others and that policies and rules of interaction should be chosen—at least in part—according to how well they achieve ethical outcomes (e.g., Bentham, 1789; Rawls, 1971; Sen, 1999). When outcomes matter, policy choices need to be made in light of human nature. What incentives encourage people to contribute to a public good, such as clean air? If people are starving and need to be fed, will collective incentive systems succeed in feeding them? If racial equality in the workplace is the goal, will this be best achieved by seminars designed to ferret out negative stereotypes in the attitudes of participants? Or will this increase hostility, making matters worse?

The nature of human nature matters for a second reason: It may place constraints on what can be considered a moral imperative. An action cannot be morally required unless it is possible to perform. But when it comes to human behavior, the meaning of *possible* is complicated (see Conclusions). Consider the following example. Corporations have many internal rules regulating procedures, a (large) subset of which are not safety rules. Yet violations of these rules can produce cascades of consequences that end up being ethically catastrophic (think Homer Simpson at the nuclear plant). Perhaps people should be alert to such violations; perhaps this should be a moral imperative, in the same way that monitoring the safety of one's child is. The mind is designed to moni-

tor for breaches of safety rules (Fiddick, Cosmides, and Tooby, 2000; Stone et al. 2002), and certain conditions, such as impending parenthood, seem to hyperactivate this system (Leckman and Mayes, 1998, 1999). But what if the human mind lacks cognitive mechanisms that spontaneously monitor for violations of procedural rules when these are not, in any obvious way, about safety? If this were true, could a person be held ethically responsible for not noticing such a breach? As we will see, this example is not as science-fictional as it may seem.

There is yet a third reason that ethicists should care about human nature: Ethicists are human beings. If the human cognitive architecture contains programs that generate moral intuitions in humans, then it generates moral intuitions in humans who are ethicists. These evolved programs cause certain moral intuitions to be triggered by particular situations. Yet this in no way justifies those moral intuitions—see above. Indeed, on reflection, some of these moral intuitions may be found wanting (yes, the ability to reflect is also made possible by evolved programs; see Leslie, 1987; Frith, 1992; Baron-Cohen, 1995; Cosmides and Tooby, 2000a). If outcomes matter to ethical judgments, then ethicists need to focus on the real world consequences of alternative policies, and not have their judgment unduly affected by moral sentiments that are nothing more than read-outs of evolved programs that were generated by an amoral process.

Justified or not, people's moral sentiments are a fact of life that anyone in business will need to accommodate. Far more needs to be known about the evolutionary psychology of moral reasoning and moral sentiments, but a start has been made. Below we present a brief overview of where evolutionary psychology fits in the intellectual landscape. Then we present empirical findings from evolutionary psychology relevant to three different topics: social exchange, collective action, and the social construction of groups. Some findings, like the results about social exchange, rest on a large evidentiary base that also includes cross-cultural tests. Others are newer, and more tentative. We offer these findings not as the last word on each topic, but as food for thought. For each topic, we briefly discuss possible implications for business ethics. Our intention is not to present well-worked out ethical theories in these sections. Instead, they are offered in the spirit of brain storming, as an exercise in how research in evolutionary psychology might eventually inform ethical theory and practice.

What is Evolutionary Psychology?

In the final pages of the *Origin of Species*, after Darwin had presented the theory of evolution by natural selection, he made a bold prediction: "In the distant future I see open fields for far more important researches. Psychology will be based on a new foundation, that of the necessary acquirement of each mental power and capacity by gradation." More than a century later, a group of scientists—Martin Daly, Margo Wilson, Don Symons, John Tooby, Leda Cosmides, David Buss, Steve Pinker, Gerd Gigerenzer—began to work out exactly how Darwin's fundamental insights could be used as a foundation on which to build

a more systematic approach to psychology (for review, see Tooby and Cosmides, 1992; see also Symons, 1979; Cosmides and Tooby, 1987; Daly and Wilson, 1988; Buss, 1989; Pinker, 1997; Gigerenzer, 2000). We were motivated by new developments from a series of different fields:

Advance #1. The cognitive revolution was providing, for the first time in human history, a precise language for describing mental mechanisms, as programs that process information.

Advance #2. Advances in paleoanthropology, hunter-gatherer studies and primatology were providing data about the adaptive problems our ancestors had to solve to survive and reproduce and the environments in which they did so.

Advance #3. Research in animal behavior, linguistics, and neuropsychology was showing that the mind is not a blank slate, passively recording the world. Organisms come factory-equipped with knowledge about the world, which allows them to learn some relationships easily, and others only with great effort, if at all. Skinner's hypothesis—that learning is a simple process governed by reward and punishment—was simply wrong.

Advance #4. Evolutionary game theory was revolutionizing evolutionary biology, placing it on a more rigorous, formal foundation of replicator dynamics. This clarified how natural selection works, what counts as an *adaptive* function, and what the criteria are for calling a trait an *adaptation*.

We thought that, if one were careful about the causal connections between these disciplines, these new developments could be pieced together into a single integrated research framework, in a way that had not been exploited before because the connections ran between fields rather than cleanly within them. We called this framework *evolutionary psychology*.[1] The goal of research in evolutionary psychology is to discover, understand, and map in detail the design of the human mind, as well as to explore the implications of these new discoveries for other fields. The eventual aim is to map *human nature*—that is, the species-typical information-processing architecture of the human brain.

Like all cognitive scientists, when evolutionary psychologists refer to "the mind," they mean the set of information-processing devices, embodied in neural tissue, that are responsible for all conscious and nonconscious mental activity, and that generate all behavior. And like other psychologists, evolutionary psychologists test hypotheses about the design of these information-processing devices—these programs—using laboratory methods from experimental cognitive and social psychology, as well as methods drawn from experimental economics, neuropsychology, and cross-cultural field work.

What allows evolutionary psychologists to go beyond traditional approaches in studying the mind is that they make active use in their research of an often overlooked fact: That the programs comprising the human mind were designed by natural selection to solve the adaptive problems faced by our hunter-gatherer ancestors—problems like finding a mate, cooperating with others,

hunting, gathering, protecting children, avoiding predators, and so on. Natural selection tends to produce programs that solve problems like these reliably, quickly, and efficiently. Knowing this allows one to approach the study of the mind like an engineer. You start with a good specification of an adaptive information-processing problem, then you do a task analysis of that problem. This allows you to see what properties a program would have to have in order to solve that problem well. This approach allows you to generate testable hypotheses about the structure of the programs that comprise the mind.

From this point of view, there are precise causal connections that link the four developments above into a coherent framework for thinking about human nature and human society (Tooby and Cosmides, 1992). These connections (C-1 through C-6) are as follows:

C-1. Each organ in the body evolved to serve a function: the intestines digest, the heart pumps blood, the liver detoxifies poisons. The brain is also an organ, and its evolved function is to extract information from the environment and use that information to generate behavior and regulate physiology. From this perspective, the brain is a computer, that is, a physical system that was designed to process information (*Advance #1*). Its programs were designed not by an engineer, but by natural selection, a causal process that retains and discards design features on the basis of how well they solved problems that affect reproduction (*Advance #4*).

The fact that the brain processes information is not an accidental side-effect of some metabolic process: The brain was designed by natural selection *to be* a computer. Therefore, if you want to describe its operation in a way that captures its evolved function, you need to think of it as composed of programs that process information. The question then becomes, what programs are to be found in the human brain? What are the reliably developing, species-typical programs that, taken together, comprise the human mind?

C-2. Individual behavior is generated by this evolved computer, in response to information that it extracts from the internal and external environment (including the social environment) (*Advance #1*). To understand an individual's behavior, therefore, you need to know both the information that the person registered *and* the structure of the programs that generated his or her behavior.

C-3. The programs that comprise the human brain were sculpted over evolutionary time by the ancestral environments and selection pressures experienced by the hunter-gatherers from whom we are descended (*Advances #2* and *#4*). Each evolved program exists because it produced behavior that promoted the survival and reproduction of our ancestors better than alternative programs that arose during human evolutionary history. Evolutionary psychologists emphasize hunter-gatherer life because the evolutionary process is slow—it takes tens of thousands of years to build a program of any complexity. The industrial revolution—even the agricultural revolution—are mere eyeblinks in evolutionary time, too short to have selected for new cognitive programs.

C-4. Although the behavior our evolved programs generate would, on average, have been adaptive (reproduction-promoting) in ancestral environments, there is no guarantee that it will be so now. Modern environments differ importantly from ancestral ones—particularly when it comes to social behavior. We no longer live in small, face-to-face societies, in semi-nomadic bands of 50-100 people, many of whom were close relatives. Yet our cognitive programs were designed for that social world.

C-5. Perhaps most importantly, the brain must be comprised of many different programs, each specialized for solving a different adaptive problem our ancestors faced—i.e., the mind cannot be a blank slate (*Advance #3*).

In fact, the same is true of any computationally powerful, multi-tasking computer. Consider the computer in your office. So many people analyze data and write prose that most computers come factory-equipped with a spreadsheet and a text-editor. These are two separate programs, each with different computational properties. This is because number-crunching and writing prose are very different problems: the design features that make a program good at data analysis are not well-suited to writing and editing articles, and vice versa. To accomplish both tasks well, the computer has two programs, each well-designed for a specific task. The more functionally specialized programs it has, the more intelligent your computer is: the more things it can do. The same is true for people.

Our hunter-gatherer ancestors were, in effect, on a camping trip that lasted a lifetime, and they had to solve many different kinds of problems well to survive and reproduce under those conditions. Design features that make a program good at choosing nutritious foods, for example, will be ill-suited for finding a fertile mate. Different problems require different evolved solutions.

This can be most clearly seen by using results from evolutionary game theory (*Advance #4*) and data about ancestral environments (*Advance #2*) to define adaptive problems, and then carefully dissecting the computational requirements of any program capable of solving those problems. So, for example, programs designed for logical reasoning would be poorly-designed for detecting cheaters in social exchange, and vice versa; as we will show, it appears that we have programs that are functionally specialized for reasoning about reciprocity and exchange.

C-6. Lastly, if you want to understand human culture and society, you need to understand these domain-specific programs. The mind is not like a video camera, passively recording the world but imparting no content of its own. Domain-specific programs organize our experiences, create our inferences, inject certain recurrent concepts and motivations into our mental life, give us our passions, and provide cross-culturally universal frames of meaning that allow us to understand the actions and intentions of others. They cause us to think certain very specific thoughts; they make certain ideas, feelings, and reactions seem reasonable, interesting, and memorable. Consequently, they play a key role in determining which ideas and customs will easily spread from mind to mind, and which will not. That is, they play a crucial role in shaping human culture.

Instincts are often thought of as the diametric opposite of reasoning. But the reasoning programs that evolutionary psychologists have been discovering (i) are complexly specialized for solving an adaptive problem; (ii) they reliably develop in all normal human beings; (iii) they develop without any conscious effort and in the absence of formal instruction; (iv) they are applied without any awareness of their underlying logic, and (v) they are distinct from more general abilities to process information or behave intelligently. In other words, they have all the hallmarks of what we usually think of as an instinct (Pinker, 1994). In fact, one can think of these specialized circuits as *reasoning instincts*. They make certain kinds of inferences just as easy, effortless and "natural" to us as humans, as spinning a web is to a spider or building a dam is to a beaver.

Consider this example from the work of Simon Baron-Cohen (1995), using the Charlie task. A child is shown a schematic face ("Charlie") surrounded by four different kinds of candy. Charlie's eyes are pointed toward the Milky Way bar (for example). The child is then asked, "Which candy does Charlie want?" Like you and I, a normal 4 year old will say that Charlie wants the Milky Way—the candy Charlie is looking at. In contrast, children with autism fail the Charlie task, producing random responses. However—and this is important—when asked which candy Charlie is looking at, children with autism answer correctly. That is, children with this developmental disorder can compute eye direction correctly, *but they cannot use that information to infer what someone wants*.

We know, spontaneously and with no mental effort, that Charlie *wants* the candy he is *looking at*. This is so obvious to us that it hardly seems to require an inference at all. It is just common sense. But "common sense" is caused: it is produced by cognitive mechanisms. To infer a mental state (*wanting*) from information about eye direction requires a computation. There is a little inference circuit—a reasoning instinct—that produces this inference. When the circuit that does this computation is broken or fails to develop, the inference cannot be made. Those with autism fail the Charlie task because they lack this reasoning instinct.

As a species, we have been blind to the existence of these instincts—not because we lack them, but precisely because they work so well. Because they process information so effortlessly and automatically, their operation disappears unnoticed into the background. These instincts structure our thought so powerfully that it can be difficult to imagine how things could be otherwise. As a result, we take normal behavior for granted: We do not realize that normal behavior needs to be explained at all.

For example, at a business school, all aspects of trade are studied. Business school scholars and students take for granted the fact that, by exchanging goods and services, people can make each other better off. But this kind of cooperation for mutual benefit—known in evolutionary biology as reciprocity, reciprocal altruism, or social exchange—is not common in the animal kingdom. Some species—humans, vampire bats, chimpanzees, baboons—engage in this very useful form of mutual help, whereas others do not (Cashdan, 1989; Isaac, 1978; Packer, 1977; de Waal, 1989; Wilkinson, 1988).

This rarity is itself telling: It means that social exchange is not generated by a simple general learning mechanism, such as classical or operant conditioning. All organisms can be classically and operantly conditioned, yet few engage in exchange. This strongly suggests that engaging in social exchange requires specific cognitive machinery, which some species have and others lack. That is, there are good reasons to think we humans have cognitive machinery that is functionally specialized for reasoning about social exchange—reasoning instincts that make thinking about and engaging in social exchange as easy and automatic for humans as stalking prey is for a lion or building a nest is for a bird.

But what, exactly, are these programs like? The research we have been conducting with our colleagues on the cognitive foundations of social exchange—of trade—suggests that the programs that allow social exchange to proceed in humans are specialized for that function, and include a subroutine that one can think of as an instinct that causes a certain kind of moral reasoning: the detection of cheaters.

The Cognitive Foundations of Trade

Selection pressures favoring social exchange exist whenever one organism (the provisioner) can change the behavior of a target organism to the provisioner's advantage by making the target's receipt of a provisioned benefit *conditional* on the target acting in a required manner. This mutual provisioning of benefits, each conditional on the others' compliance, is what is meant by social exchange or reciprocation (Cosmides, 1985; Cosmides and Tooby, 1989; Tooby and Cosmides, 1996). Social exchange is an "I'll scratch your back if you scratch mine" principle: *X* provides a benefit to *Y* conditional on *Y* doing something that *X* wants.

Robert Trivers, W. D. Hamilton, Robert Axelrod, and other evolutionary researchers used game theory to understand the conditions under which social exchange can and cannot evolve (Trivers, 1971; Axelrod and Hamilton, 1981; Boyd, 1988). For adaptations causing this form of cooperation to evolve and persist—that is, for reciprocation to be an evolutionarily stable strategy (ESS)—the behavior of cooperators must be generated by programs that perform certain specific tasks well. For example these programs would need design features that would (i) allow cooperators to detect cheaters (i.e., those who do not comply or reciprocate), and (ii) cause cooperators to channel future benefits to reciprocators, not cheaters (Trivers, 1971; Axelrod and Hamilton, 1981; Axelrod, 1984).

In other words, reciprocation cannot evolve if the organism lacks reasoning procedures that can effectively detect cheaters (i.e., those who take conditionally offered benefits without providing the promised return). Such individuals would be open to exploitation, and hence selected out. Based on such analyses, Cosmides and Tooby hypothesized that the human neurocognitive architecture includes *social contract algorithms*: a set of programs that were specialized by natural selection for solving the intricate computational problems inherent in adaptively engaging in social exchange behavior, including a subroutine for cheater detection.

Conditional Reasoning

Reciprocation is, by definition, social behavior that is conditional: you agree to deliver a benefit *conditionally* (conditional on the other person doing what you required in return). Understanding it therefore requires conditional reasoning.

Indeed, an agreement to exchange—a social contract—can be expressed as a conditional rule: ***If*** *A provides a requested benefit to or meets the requirement of B,* ***then*** *B will provide a rationed benefit to A.* A cheater is someone who illicitly takes the benefit specified in the social contract; that is, someone who violates the social contract by taking the benefit without meeting the provisioner's requirement.

Because engaging in social exchange requires conditional reasoning, investigations of conditional reasoning can be used to test for the presence of social contract algorithms. The hypothesis that the brain contains social contract algorithms predicts a sharply enhanced ability to reason adaptively about conditional rules when those rules specify a social exchange. The null hypothesis is that there is nothing specialized in the brain for social exchange: This predicts no enhanced conditional reasoning performance specifically triggered by social exchanges as compared to other contents.

A standard tool for investigating conditional reasoning is Wason's 4-Card Selection Task (Wason, 1966, 1983; Wason and Johnson-Laird, 1972). Using this task, Cosmides, Tooby, and their colleagues conducted an extensive series of experiments to address the following questions:

1. Do our minds include cognitive machinery that is *specialized* for reasoning about social exchange? (alongside some other domain-specific mechanisms, each specialized for reasoning about a different adaptive domain involving conditional behavior?) Or,
2. Is the cognitive machinery that causes good conditional reasoning general—does it operate well regardless of content? (a blank slate-type theory; Pinker, 2002).

This second, blank-slate view was in trouble before we even started our investigations. If the human brain had cognitive machinery that causes good conditional reasoning regardless of content, then people should be good at tasks requiring conditional reasoning. For example, they should be good at detecting violations of conditional rules. Yet studies with the Wason selection task had already shown that they are not. The Wason task asks you to look for potential violations of a conditional rule (*If P then Q*), such as "If a person has Ebbinghaus disease, then that person is forgetful" (see Figure 1 [p. 100], panel a). The rule is accompanied by pictures of four cards, each representing one person—a patient in this case. For each card, one side tells whether the patient in question has Ebbinghaus disease, and the other side tells whether that patient is forgetful. However, you can see only one side of each card, so your information about each patient is incomplete. The question: Which card(s) would you need to turn over to find out if there are patients whose situation violates the rule?

Figure 1

a. General Structure of a Descriptive Problem

Consider the following rule: If P then Q.
The cards below have information about four situations. Each card represents one situation. One side of a card tells whether *P* happened, and the other side of the card tells whether *Q* happened. Indicate only those card(s) you definitely need to turn over to see if any of these situations violate the rule.

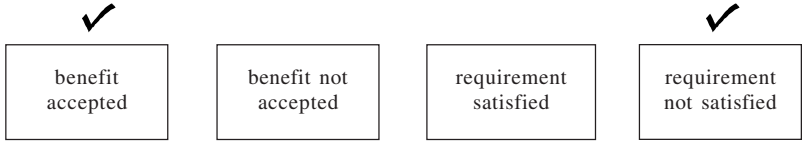
b. General Structure of a Social Contract Problem

Consider the following rule:

standard form:
If you take the *benefit*, then you satisfy the *requirement*.

switched form:
If you satisfy the *requirement*, then you take the *benefit*.
If P then Q

The cards below have information about four people. Each card represents one person. One side of a card tells whether a person accepted the benefit, and the other side of the card tells whether that person satisfied the requirement. Indicate only those card(s) you definitely need to turn over to see if any of these people are violating the rule.



standard:	*P*	*not-P*	*Q*	*not-Q*
switched:	*Q*	*not-Q*	*P*	*not-P*

Legend for Figure 1. The Wason selection task. The conditional rule (If *P* then *Q*) always has specific content. *Panel A.* The general structure of the task in logical terms. Check marks indicate the logically correct card choices. *Panel B.* The general structure of the task when the content of the conditional rule expresses a social contract. It can be translated into either logical terms (*P*s and *Q*s) or social contract terms (benefits and requirements). Here, check marks indicate the correct card choices if one is looking for cheaters: the *benefit accepted* card and the *requirement not satisfied* card (regardless of the logical category to which these correspond). For example, (i) "If you give me your watch, I'll give you $100" and (ii) "If I give you $100, then you give me your watch" express the same offer to exchange—the same social contract. Standard form is where the benefit to the potential cheater appears in the antecedent clause (*P*); switched is where the benefit appears in the consequent clause (*Q*). Thus, if I were the potential cheater, then (i) is standard form [because the benefit to me (getting your watch) appears in the antecedent clause, *P*)] and (ii) is switched form. In either case, my taking your watch without giving you the promised payment would count as cheating. Whereas these cards fall into the logical categories *P & not-Q* for a rule expressed in standard form, they fall into the logical categories *Q and not-P* for a rule expressed in switched form. *Q & not-P* is not a logically correct response. It is, however, the adaptively correct, cheater detection response when the benefit is in the consequent clause.

One card says "has Ebbinghaus disease" (i.e., *P*), one says "does not have Ebbinghaus disease" (*not-P*), one says "is forgetful" (Q), and one says "is not forgetful" (*not-Q*).

A conditional rule like this is violated whenever *P* happens but *Q* does not happen (in this case, whenever someone has Ebbinghaus disease but is not forgetful). To respond correctly, you would need to check the patient who has Ebbinghaus disease and the patient who is not forgetful (i.e., *P & not-Q*). Yet studies in many nations have shown that reasoning performance on descriptive rules like this is low: only five to thirty percent of people give the logically correct answer, even when the rule involves familiar terms drawn from everyday life[2] (Cosmides, 1989; Wason, 1966, 1983; Wason and Johnson-Laird, 1972).

Are people also poor at detecting cheaters? To show that people who do not spontaneously look for violations of conditional rules can do so easily when the conditional rule expresses a social contract and a violation represents cheating would be (initial) evidence that the mind has reasoning procedures specialized for detecting cheaters.

That is precisely the pattern found. People who ordinarily cannot detect violations of if-then rules can do so easily and accurately when that violation represents cheating in a situation of social exchange. Given a conditional rule of the general form, "If you take benefit B, then you must satisfy requirement R" (e.g., "If you borrow my car, then fill up the tank with gas"), people will check the person who accepted the benefit (borrowed the car; *P*) and the person who did not satisfy the requirement (did not fill the tank; *not-Q*)—the individuals that represent potential cheaters (see Figure 1, panel b, standard form). The adaptively correct answer is immediately obvious to almost all subjects, who commonly experience a pop-out effect. No formal training is needed. Whenever the content of a problem asks one to look for cheaters in a social exchange, subjects experience the problem as simple to solve, and their performance jumps dramatically. In general, sixty-five to eighty percent of subjects get it right, the highest performance found for a task of this kind (for reviews, see Cosmides and Tooby, 1992, 1997, 2000b; Fiddick, Cosmides, and Tooby, 2000; Gigerenzer, 1992; Platt and Griggs, 1993).

It is not familiarity. This good performance has nothing to do with the familiarity of the rule tested. First, familiarity does not enhance performance for descriptive rules. Second (and most surprising), people are just as good at detecting cheaters on culturally unfamiliar or imaginary social contracts as they are for ones that are completely familiar, providing a challenge for any counterhypothesis resting on a general-learning skill acquisition account. An unfamiliar, culturally alien rule—e.g., "If a man eats cassava root, then he must have a tattoo on his face"—can elicit excellent cheater detection. All one needs to do is embed it in a scenario that says that the people involved consider eating cassava root to be a benefit; the rule then implies that having a tattoo is the requirement one must satisfy to be eligible for that benefit (Cosmides, 1985, 1989; Gigerenzer and Hug, 1992; Platt and Griggs, 1993).

It is not logic. Further experiments showed that subjects do not choose *P & not-Q* on social contract problems because these problems activate *logical* reasoning. Instead, they activate a differently patterned, specialized, logic of social exchange (Cosmides and Tooby, 1989).

Formal logic (i.e., the propositional calculus) is content-independent: a logical violation has occurred whenever *P* happens and *Q* does not happen. In contrast, cheater detection requires the search for illicitly taken *benefits.* It does not matter whether this benefit is in the antecedent clause (*P*) or the consequent clause (*Q*): regardless of logical category, one must check the person who accepted the benefit and the person who did not meet the requirement. It is possible to construct a social exchange problem for which formal logic and social exchange logic predict different answers. When this is done, subjects overwhelmingly follow the evolved logic of social exchange. They investigate anyone who has taken the benefit and anyone who has not satisfied the requirement that it was contingent upon, even if this results in a logically incorrect answer, such as *Q and not-P*. (See Figure 1, panel b, on switched social contracts; Cosmides, 1985, 1989; Gigerenzer and Hug, 1992).

It is not a general ability to reason about permission rules (not a deontic logic). Detecting cheaters on social contracts was an important adaptive problem for our ancestors; so was detecting when people are in danger because they have failed to take appropriate precautions (Fiddick, Cosmides, and Tooby, 2000). Experimental results show that people are good at detecting violations of these two classes of conditional rules (precautions have the general form, "*If one is to engage in hazardous activity H, then one must take precaution R*"; Fiddick, Cosmides, and Tooby, 2000; Stone et al., 2002). Note, however, that social contracts and precautionary rules are instances of a more general class, *permission rules*. A permission rule is a conditional rule specifying the conditions under which one is permitted to take an action. They have the form "If action A is to be taken, then precondition C must be met" (Cheng and Holyoak, 1985). There are, however, permission rules that are neither social contracts nor precautionary rules. Indeed, we encounter many rules like this in everyday life—bureaucratic or corporate rules often state a procedure that is to be followed without specifying a benefit (or a danger). Despite their ubiquity in modern life, people are not good at detecting violations of permission rules when these are neither social contracts nor precautionary rules (Cosmides and Tooby, 1992; Manktelow and Over, 1991; Barrett, 1999; see below).

The Design of Social Exchange Mechanisms

Many cognitive scientists have now investigated social contract reasoning, and many of the predicted design features have been tested for and found. For example:

1. One needs to understand each new opportunity to exchange as it arises, so social exchange reasoning should operate even for unfamiliar social

contract rules. That is the case: Cheater detection occurs even when the social contract is wildly unfamiliar (see above).

2. The mind's automatically deployed definition of cheating is tied to the perspective one is taking (Gigerenzer and Hug, 1992). Given the rule "If an employee is to get a pension, then that employee must have worked for the firm for over 10 years," different answers are given depending on whether subjects are cued into the role of employer or employee. The former look for cheating by employees, investigating cases of *P* and *not-Q* (employees with pensions; employees who have worked for fewer than 10 years); the latter look for cheating by employers, investigating cases of *not-P* and *Q* (employees with no pension; employees who have worked more than 10 years).
3. To elicit cheater detection, the rule must specify a benefit: if there is no benefit, then it is not a social contract. Consider, for example, the following two rules granting conditional permission: (i) "If you are to go out at night, then you must tie a small rock to your ankle" versus (ii) "If you are to take out the garbage at night, then you must tie a small rock to your ankle." For our subject population, going out at night is seen as a benefit, but taking out the garbage is not. People succeed at (i) but not at (ii), even though both are rules granting conditional permission (of the form, "If you are to take *action A*, then you must satisfy *requirement R*). When *action A* of a permission rule is difficult to interpret as a conditional benefit (as in ii),[3] people have trouble detecting violations (Cosmides and Tooby, 1992; Manktelow and Over, 1991; Barrett, 1999).
4. For the reasoning enhancement to occur, the violations must potentially reveal cheaters: individuals who violate the rule *intentionally*. When detecting violations of social contracts would reveal only innocent mistakes, enhancement does not occur (Barrett, 1999; Cosmides and Tooby, 2000b; Fiddick, 1998, in press).
5. Excellence at cheater detection is not a skill elicited by extensive participation in an advanced market economy. Consistent with its being a species-typical ability, social contract reasoning effects are found across cultures, from industrial democracies to hunter-horticulturalist groups in the Ecuadorian Amazon (Sugiyama, Tooby, and Cosmides, 2002).

Perhaps the strongest evidence that there is a neural specialization designed for cheater detection is the discovery that cheater detection can be selectively impaired by brain damage, without impairing other reasoning abilities. R.M., a patient with extensive brain damage, was given a large battery of Wason tasks that were formally identical and matched for difficulty. His ability to detect violations of precautionary rules was very good, but his ability to detect cheaters on social contracts was very impaired (Stone et al. 2002). If performance on social contract and precautionary rules were a byproduct of some more general ability to reason, then damage to that more general mechanism would impair reasoning on both types of problem—not just on social contract problems.

These findings are all direct predictions of the hypothesis that there are neurocognitive mechanisms specialized for reasoning about social exchange. They are not predicted by other views. Alternative hypotheses to explain reasoning on the Wason selection task have been proposed (e.g., Cheng and Holyoak, 1985; Manktelow and Over, 1991; Sperber, Cara, and Girotto, 1995), but none so far can account for this array of results (for reviews, see Cosmides and Tooby, 1992, 1997, 2000b; Fiddick, Cosmides, and Tooby, 2000).

Cognition, institutions, and culture. An advanced market economy would be impossible for a species lacking social contract algorithms. Although necessary, possession of these cognitive programs is not a sufficient condition for the emergence of a market economy. Other institutions (i.e., "rules of the game"; North, 1990) and sociocultural conditions need to be co-present. Examples include a wide, consensually shared numerical counting system, a culturally accepted medium of exchange (which reduces transaction costs compared to barter), a division of labor making exchange more profitable, wide-broadcast information systems that signal scarcity of resources (e.g., prices), and institutions that enforce contracts and punish cheaters (e.g., rule of law). When these modern cultural institutions are combined with social contract algorithms designed for an ancestral world, what should we expect?

Implications for Understanding Business and Business Ethics

Cheating on an agreement to trade is obviously an ethical violation. The research discussed above suggests that our minds are well-designed for detecting this form of ethical violation. However, the cognitive machinery that does this was designed for small-scale, face-to-face societies: hunter-gatherer bands in which you lived in close contact with the same group of individuals day in and day out, and where it was relatively easy to see whether the conditions of a social contract had been met. Detecting cheaters in a large market economy may be far more difficult.

Too many people for the mind to process. Modern corporations employ hundreds or thousands of people, most of whom are strangers to one another. The working interactions of these individuals is governed by implicit and explicit agreements to provide services to one another, to the corporation, to suppliers, clients, and customers. But the more people are involved, the more difficult detecting cheaters should be. There are two reasons for this.

The first is cognitive load. More interactants and interactions means more opportunities to cheat, each of which needs to be monitored. Data from modern hunter-gatherers and hunter-horticulturalists show that individuals usually limit themselves to a rather small number of regular exchange partners within their larger group (Gurven, 2002; Gurven et al. 2000; modern hunter-gatherer bands average 50 people, including children (Lee and DeVore, 1968). A memory system designed for this level of interaction would need separate slots for each interactant, each slot able to store a certain amount of data about one's history of interaction with that person (Cosmides and Tooby, 1989). But how many

person-slots is the memory system equipped with, and how much information can each hold (and for how long)? No one knows the answer to these questions yet. But monitoring the large number of persons and interactions in a modern corporation might overload the system, especially if these person-slots turn out to be limited in their number and capacity to hold information.

The second difficulty arises from lack of transparency. Hunter-gatherers have little privacy: if the neighbor my family has been helping comes back from a hunt with lots of meat, this will be seen and everyone will know, making nonreciprocation socially difficult for him. But when many people are interacting in complex ways, it can be difficult to tell when someone is failing to fully satisfy the responsibilities for which they are paid a salary. Has an employee generated enough ideas for ad campaigns or succeeded at improving public relations? Some tasks require the production of abstract goods that are difficult to quantify. Are profits really that high and debt that low? Complex numerical accounting systems are necessary in a market economy, but hardly transparent.

These problems can be mitigated by creating small internal working groups (on a more hunter-gathererish scale), and by creating more transparent and public systems for keeping track of who has done what for whom. Both are obvious solutions that many companies have converged on.

No design for detecting procedural violations. Cheater detection is most strongly activated when two conditions are jointly met: (i) the rule specifies a contingent *benefit* that a potential violator could illicitly obtain, and (ii) one suspects the rule will be violated *intentionally*. There is, of course, a connection between these two: the prospect of gaining a benefit at no cost provides an incentive to cheat that may tempt one to intentionally violate a social contract.

In this light, consider these ordinary procedural rules that one might encounter in a business:

(i) "If the invoice is from the appliances department, then route it through accounting division B."

(ii) "If you are going to contact a disgruntled client, first notify the manager."

Both rules are deontic (i.e., rules prescribing the conditions under which one is entitled or obligated to do something). But neither is a social contract because neither specifies a *benefit* that one is entitled to only if a requirement has been met (the pleasures of routing appliance invoices are difficult to fathom, and most people dread speaking to disgruntled clients). Moreover, such rules are more likely to be broken through inattention or negligence, rather than on purpose (what is there to gain?). When rules do not regulate access to benefits and violations are likely to be mistakes, cheater detection is barely activated; under these circumstances, fewer than thirty percent of people detect violations (Barrett, 1999; Cosmides and Tooby, 2000b).

Yet, depending on the downstream consequences, failure to comply with rules like these may ultimately waste resources or even endanger consumers. In many cases, the public—and the business itself—would be better served if such violations were detected. This is more likely to happen when people's

attention is spontaneously drawn to situations that might involve violations, and the research discussed above shows that this is more likely to happen when rules are restated or reframed as social contracts (or else as precautionary rules, see above). For example, rule (i) could be prefaced with the explanation that the ovens, stoves and dishwashers in the appliance department are very expensive, high profit items, and division B specializes in making sure consumers make payments on these items on time. This background enables one to reframe the rule as, "If the consumer is to buy an expensive item from us, then they must pay for it in a timely manner"—a classic social contract. Rule (ii) could be prefaced with an explanation that dealing with disgruntled customers can be difficult or, if they are litigious, hazardous to the company, and notifying the manager is a precautionary measure taken because the manager may have some timely advice to give. This background reframes the rule as a classic precautionary rule, one of the other domains for which our minds are designed to detect violations.

The agency problem can bedevil attempts to reframe. A benefit to the company is not necessarily a benefit to one of the company's agents; agents who do not see the company's costs and benefits as such are unlikely to spontaneously attend to potential cases of cheating by employees or clients. In some cases, reframing the action specified in the rule as a benefit or hazard may be difficult, for example, when the rationale for the rule is too obscure or when the benefits to be gained are too many steps removed from the action that the rule permits or obliges.

It is important that people be alert to the possibility of rule violations when these might have negative ethical consequences. The best way to do this is for companies to work with human nature, rather than against it. The more procedural rules can be reframed as social contracts or precautions, the more one will engage employees' spontaneous attention to potential rule violations. To the extent such reframings are possible, ethicists advising businesses might want to suggest them.

Moral Sentiments: The Desire to Punish Free Riders on Collective Actions

Dyadic cooperation is sometimes seen in the animal kingdom. Far rarer are collective actions: cooperation between three or more individuals to achieve a common goal. Yet this form of multi-individual cooperation is common in our own species. It occurs not only in modern circumstances, but in hunter-gatherer and hunter-horticulturalist societies as well. Common examples include intergroup conflict—band-level warfare—cooperative hunting, and certain community-wide projects, such as shelter-building.

In these circumstances, sets of individuals cooperate to achieve a common goal, and they do so even when that goal is a public good—that is, even when the rewards to individuals are not intrinsically linked to individual effort. This has been, and continues to be, a puzzle to both economists and evolutionary biologists. When faced with the decision to participate in a collective action, there are two choices: free ride or participate. Ever since Mancur Olson's trenchant analysis, rational choice theorists have understood that free riding generates a higher payoff than cooperation: Participants and free riders get the same benefit—a successful outcome—but free riders do not incur the cost of participation (Olson, 1965). This incentive to free ride results in a paradoxical outcome: Participation unravels and the project fails, even though each individual would have been better off if the project's goal had been successfully achieved.

Evolutionary biologists find cooperation in collective actions puzzling for a different, but related, reason. In evolutionary biology, the different payoffs to alternative choices are relevant only if they cause differential reproduction of alternative designs (alternative programs) that cause those choices. The fact that collective action is rare in the animal kingdom means that most organisms *lack* programs that cause participation: free riding, therefore, is the default choice. If payoffs to collective action translate into reproductive advantages, then how could designs causing participation have gained a toe-hold in a universe dominated by non-participants? Those who participated in a successful collective action would have experienced an increase in their fitness, but free riders would have benefited even more (by getting the benefits of the achieved goal without suffering the costs of participation). The currency is differential reproduction of participant- versus free-riding designs; this means that individuals equipped with programs that caused free-riding would have out-reproduced those equipped with programs that caused participation. Consequently, free-rider designs would have been selected for, and any participation-designs that arose in a population would have been selected out. If so, then why do we see individual human beings routinely and willingly participating in collective actions? Is this a byproduct of adaptations that evolved for some other purpose, or did evolution produce mechanisms designed to cause this form of cooperation?

There may not be adaptations designed for regulating participation in collective actions. But if there are, programs that cause participation would need to be equipped with strategies that eliminated the fitness advantage of free riders. Without such features, designs causing participation could not be evolutionarily stable strategies (Maynard Smith, 1982). Price, Cosmides, and Tooby (2002) have proposed that punitive sentiments toward free riders are generated by an adaptation in participant designs whose function is to eliminate the fitness advantage free rider designs would otherwise enjoy. They tested this hypothesis against a labor recruitment theory and rational choice theory.

Alternative Theories of Adaptive Function of a Moral Sentiment

All functional theories, evolutionary or economic, propose that one's willingness to participate in a collective action will be a function of how much one expects to individually benefit from its success. But theories diverge in their predictions about the conditions that should trigger punitive sentiments toward free riders (as well as the conditions that should trigger pro-reward sentiments toward participants).

The adaptive function of a program is the reason it evolved: the selective advantage that, over evolutionary history, caused the program in question to be favored over alternative ones. If *eliminating free rider fitness advantages* were the adaptive function of punitive sentiments toward free riders, then several predictions (E-1 through E-6) follow about the design of the motivational system that triggers them:

E-1. An individual's own participation should be the specific factor that triggers punitive sentiments toward free riders. This is because (ancestrally) *only those individuals who contributed were at risk* of incurring lower fitness relative to free riders.

E-2. The more an individual contributes, the greater the adverse fitness differential s/he potentially suffers relative to free riders. Hence a sentiment designed to prevent outcompetition by free riders should key the *degree* of punitive sentiment toward free riders to the individual's own willingness to participate: The more one participates, the more punitive one should feel toward free riders.

E-3. Those who have an interest in the goal being achieved should be more willing to participate. However, punitive sentiment should track willingness to participate, even after controlling for self-interest in the group goal.

Indeed, if eliminating the free rider's fitness advantage were the adaptation's only function, then:

E-4. After controlling for willingness to participate, any relationship between perceived benefit and punitive sentiment should disappear.

E-5. Willingness to participate should predict punishment, but not sentiments in favor of rewarding participants. (When reward induces a free riding underproducer to join a collective action, this preserves the underproducer's relative fitness advantage compared to the producer design that is doing the rewarding).

E-6. Consequently, pro-reward sentiments should not track punitive sentiments, especially among those most willing to participate.

The *labor recruitment theory* is an alternative hypothesis about the adaptive function of punitive sentiments toward free riders on collective actions. According to this hypothesis, punitive sentiments were designed by evolution to encourage more participation in a collective action, in an effort to increase the

probability that the common goal is successfully achieved. This hypothesis leads to many of the same predictions as rational choice theory, to wit:

L-1. Those most likely to benefit from achievement of a group goal should differentially act to induce others to participate. Self-interest in the group goal should trigger punitive sentiments, and the greater one's self-interest in that goal, the more punitive one should feel toward free riders.

L-2. Self-interest should independently predict punitive sentiment (even after controlling for willingness to participate). Encouraging self-sacrifice by others provides the largest net benefit—even for a free rider.

If encouraging participation by others were the adaptation's only function, then:

L-3. After controlling for self-interest, there should be no relationship between willingness to participate and punitive sentiment.

L-4. Pro-reward sentiment should track punitive sentiment. (Nothing in the problem of labor recruitment privileges the carrot over the stick as a means of inducing participation.)

L-5. Punitive sentiment should be sensitive only to labor needs, not to free riders per se. Once manpower needs are met, the system should be indifferent to the prospering of free riders.

L-6. The system should be indifferent to whether a non-participant is a free rider or not. Self-interest in the group goal should trigger punitive sentiment toward *any* non-participant who could help achieve the goal by participating, including people who do not benefit from the collective action and people who are considered exempt (e.g., women in warfare).

L-7. Those who contribute anything less than the optimal amount should be targets of punishment (even if they are contributing at the same level as everyone else).

Price et al. (2002) compared these predictions to results from experimental economics games (e.g., Fehr and Gächter, 2000a,b; Yamagishi, 1986), and to results of a survey they conducted assessing attitudes toward participation in a collective action. This survey (which was conducted prior to September 11, 2001) asked subjects to imagine that the United States was mobilizing for war, and to indicate how strongly they agreed or disagreed with a number of statements. In addition to other variables, subjects were asked how willing they would be to participate ("If I got drafted for this war, I would probably agree to serve"), how much they felt they would benefit from the group goal being achieved ("If the USA won this war, it would be very good for me as an individual"), how punitive they would feel toward nonparticipants ("If a U.S. citizen resisted this draft, I'd think they should be punished"), and how much they felt participants should be rewarded ("If a drafted U.S. citizen agreed to serve in this war, I'd think they should be rewarded").

The Adaptive Function of Punitive Sentiments: What Do the Results Say?

The survey results were surprisingly clear cut. They supported all the predictions that follow from the hypothesis that punitive sentiments evolved to eliminate the fitness advantage that would accrue to a free-rider design (E-1 through E-6). Moreover, the results contradicted all the predictions of the labor recruitment hypothesis that they could address (L-1 through L-4). The other predictions, L-5 through L-7, were contradicted by results from public goods games in experimental economics (Fehr and Gächter, 2000a,b; see Price et al for discussion).

In short, willingness to participate was the specific trigger for punitive sentiments toward free riders: the more willing one was to participate, the more punitive one felt toward free riders. Willingness to participate independently predicted punitive sentiment, even after controlling for self-interest in the group goal (partial r = .55, .62, for two different scenarios). In contrast, self-interest in the group goal did not independently predict punitive sentiment, once the effects of willingness to participate were statistically removed.

Engineering criteria are used to recognize adaptations and deduce their functions. To discover the function of a system, one looks for evidence of special design—a design that achieves an adaptive function precisely, reliably, and economically. The motivational system that generates punitive sentiments toward free riders showed evidence of special design for eliminating the fitness advantages of free riders. For example:

1. The participation-punishment link was *selective*. Willingness to participate predicted punitive sentiment, not pro-reward sentiment.
2. The trigger for the punitive response was *precise*: Willingness to participate was the only variable to independently predict punitive sentiment. Punitiveness was not independently predicted by self-interest in the group goal or by various demographic variables.
3. The punitive response was *specific*: Willingness to participate predicted punitive sentiment toward *free riders*; once this effect was controlled for, it did not predict punitiveness more generally (toward criminals, for example).
4. The punitive response was *uniform*: The participation-punishment link was just as strong in women as in men, despite the fact that women are considered exempt from the military draft.

The empirical data from this and the public goods games contradict the predictions of rational choice theory (see Table 1). Recently, Price replicated the selectivity, precision, and specificity of the participation-punishment link using behavioral data in a totally different circumstance: an economic collective action. The subjects were Shuar hunter-horticulturalists in the Ecuadorian Amazon, a group of men participating in a collective action to cultivate and sell a sugar cane crop (Price, Barrett, and Hagen, under review). This study produced the same results as the American survey. E-1 through E-6 were supported with about the same effect sizes, providing further evidence that the motivational system that generates punitive sentiments toward free riders on collective

Table 1. Rational Choice Theory (RCT) and Moral Sentiments Toward Free Riders: Predictions versus Results

1. **RCT:** People should not punish when costs of doing so cannot be recouped. **But they do.** (e.g., Fehr and Gächter, 2000a,b)
2. **RCT:** Targets should be people who could increase their own level of cooperation to the benefit of the rational agent. So punish anyone who contributes *less than the optimum* (even if they contributed at the group average). **Yet such people are *not* punished.** (Fehr and Gächter, 2000a,b)
3. **RCT:** Self-interest in group goal *should* predict punitive sentiment. **But it does not.** (Price, Cosmides, and Tooby, 2002)
4. **RCT:** Willingness to participate should not trigger punitive sentiment *independent* of expected gain (sunk cost fallacy). **But it does.** (Price et al., 2002)
5. **RCT:** Reward should track punitive sentiment. **But it does not.** (Price et al., 2002)

Perhaps rational choice leads you to support group norms that are in your interest. But . . .

6. **RCT:** Self-interest in group goal plus willingness to participate should trigger punitive sentiment only when both are high (to avoid advocating your own punishment). **But this is not the case.** (Punitive sentiment is triggered by willingness to participate, regardless of self-interest; Price et al., 2002.)
7. **RCT:** Those who are exempt are free to punish, so there should be no willingness-punitive sentiment link in those who are exempt (e.g., women). **Yet there is.** (Price et al., 2002)
8. **RCT:** Those who are willing to participate should advocate rewarding participants (they would get the reward!). **But willingness does *not* predict pro-reward sentiment.** (Price et al., 2002)

actions was designed by natural selection to eliminate the fitness advantage that free riders would otherwise enjoy. Additionally, the Ecuadorian study, like the American one, produced evidence that directly contradicts the labor recruitment and rational choice theories.

Carrots and Sticks are Not Fungible

The claim that punitive sentiments did not evolve to solve labor recruitment problems does not imply that the human mind lacks *any* programs designed to address this problem. Indeed, Price's data suggested the presence of a motivational system designed to encourage participation: one that generates sentiments in favor of *rewarding* participants. Pro-reward sentiments were independently predicted by self-interest in the group goal. The trigger for reward sentiments was precise (only one variable, self-interest in achievement of the goal, independently predicted them) and the response triggered was specific (self-interest predicted only reward sentiments, not punitive ones).

Thus, in collective actions, the motivation to punish and the motivation to reward appear to be triggered by different variables and generated by two different systems. Most economic analyses treat reward and punishment as fungible, mere increases and decreases in utility. But this dissociation between punitive and pro-reward sentiments suggests that the carrot and the stick are not just two sides of the same coin. In a collective action, the desire to use the carrot is triggered by different circumstances than the desire to use the stick.

When Punishment is Not Possible

Ancestrally (as now), punishment is not always an option. When this is so, a participant design can avoid outcompetition by free riders if it is equipped with a feature that monitors for the presence of under-contributors and drops its own level of participation when they are present. Research on contributions to public goods in experimental economics shows that people continuously monitor the state of play, adjusting their behavior accordingly (Fehr and Gächter, 2000a,b; Kurzban, McCabe, et al., 2001). If they can, they inflict punishment on under-contributors right away (which has the secondary consequence of allowing levels of cooperation to spiral up toward the welfare-maximizing optimum of 100 percent contribution to the common pool; see Price et al., for analysis). When there is no opportunity to punish, they ratchet back their own contribution to something like the average level. As this monitoring and adjustment process iterates, contributions gradually diminish to rational choice theory expectations (Kurzban, McCabe, et al., 2001). But this iterative ratcheting back does not reflect the emergence, through learning, of rational choice: when a new collective action begins, the very same people start out contributing to the common pool at relatively high levels (about sixty percent of their endowment; rational choice theory predicts zero percent).

Implications for Business Ethics

The first thing to note is that, if the above results hold up, they indicate that punitive sentiments toward free riders in collective actions evolved for a function that, from an economic point of view, makes no sense: eliminating the fitness differential between participant-designs and free-rider designs in a distant, ancestral past. Prior to the evolution of cognitive adaptations favoring participation in collective action, no minds were designed for participation (by definition). Without adaptations promoting participation, non-participation would have been the default strategy: free riding would have been the state of nature.

But that was then, and this is now. If selection did favor adaptations for participating in collective actions, there are strong reasons to assume that, by now, these would be universal and species typical (Tooby and Cosmides, 1990). When free riding occurs now, it probably reflects a contingent choice strategy that everyone has, embedded within a set of programs that would make any of us participate in a collective action under circumstances that were more indi-

vidually auspicious. Designs that cause nothing but free riding—the original selection pressure causing participants to evolve punitive sentiments toward free riders—may no longer exist in the human population.

Punishment of free riders may be "irrational" in the rational choice sense of "rational," especially in a modern world full of anonymous strangers. But that does not matter if your goal is to understand behavior. Our evolved psychology may have been designed for a vanished world, but it generates our behavior nonetheless. People remain more afraid of spiders and snakes—ancestral dangers—than of cars and electric outlets, even though the latter pose a greater threat in the modern world. Whether it is sensible now or not, our psychology is designed so that the more we contribute to a collective action, the more punitive we will feel toward those we perceive as free riders.

Adaptations for a small world. Adaptations for participating in collective actions evolved in the context of a small social world of perhaps 20–100 people, many of whom were relatives (natural selection easily allows the evolution of mechanisms for delivering benefits to relatives noncontingently; Hamilton, 1964). As Mancur Olson pointed out in the context of labor unions, voluntary collective actions are more likely to succeed when they involve small groups rather than large ones—not surprising to an evolutionary psychologist. This occurs, he argued, because there are many compensatory benefits for those who join such groups, not merely the benefit to be gained from achieving the collective goal. His descriptions of these activities are strongly reminiscent of the risk pooling and mutual aid one sees in hunter-gatherer bands of similar size.

Large consumer and environmental advocacy groups are engaged in collective action projects, the intent of which is to curb what are seen as ethical violations by business. But large collective actions, Olson pointed out, are more difficult to sustain without the use of coercive force. An evolved psychology as described by Price *et al.* has additional implications for large collective action groups like these, especially for the morale and political attitudes of volunteers.

Morale. Idealistic people eagerly anticipate working toward noble goals with public advocacy groups. Nevertheless, many volunteers (and even paid workers) are lost to "burn-out": a catastrophic drop in morale triggered by the perception that one is doing all the work while most people free ride (often accompanied by bitterness—punitive sentiment?—toward non-participants, who are disparaged as "apathetic" or worse). The very experience of working hard for a collective good should trigger negative sentiments toward those who are not "involved." The loss of interest in making further contributions is also expected: These are private groups that lack the ability to punish free riders, a circumstance that triggers the iterative ratcheting back strategy.

Political attitudes. Less obviously, the two motivational systems—punitive sentiments triggered by degree of participation versus pro-reward sentiments triggered by self-interest in the group goal—might color the political solutions favored by various groups. For example:

Producing cleaner air is a classic public good. In an effort to reduce air pollution, one could advocate a pro-reward policy (e.g., tax incentives for businesses that contribute to the goal by reducing their pollution) or a punitive policy (e.g., fines levied on businesses that do not reduce their pollution). Which is more effective is an empirical matter, and the goal of clean air is best served by choosing the most effective policy. (N.B.: the authors have no opinion about which is best). But the very act of participating in a collective action triggers punitive sentiments toward free riders (businesses that do not reduce their pollution), not pro-reward sentiments toward participants (businesses that do reduce their pollution). Indeed, the more energetically one works for an environmental advocacy group, the more punitive one should feel toward businesses who do not curtail their pollution and toward fellow citizens who do not contribute to the group's work. Once this moral sentiment is activated, policies that impose sanctions and laws that mandate contributions toward the goal (through taxes and state agencies) may seem more reasonable and just. Indeed, individuals who, before joining an environmental advocacy group, had favored pro-reward policies might have a change of heart after joining. Once they are actively participating, they can be expected to experience an ethical tug in the direction of punitive sanctions and enforced contributions, and away from policies that reward businesses for curtailing pollution.

Working with human nature. Are there ways of harnessing these moral sentiments in the service of reducing negative externalities such as pollution? Clean air is a public good, but the individuals charged with enforcing pollution standards are government bureaucrats at agencies like the EPA, who have nothing in particular to gain by enforcement—not even the pleasure of cleaner air, if they live far from the polluters (see agency problem, above). Imagine a slightly different system: "pollution courts," where companies that had contributed to the public good by demonstrably reducing their own pollution levels had standing to both present evidence of pollution by their free-riding competitors and request the imposition of fines. Might this give companies an incentive to (i) prove they deserve standing (by lowering their own pollution levels), and (ii) investigate cases of pollution, thereby reducing the EPA's burden? Could this system wipe out the profit advantage the free riding polluter has over companies that voluntarily curtail their pollution?

Ethics and the organization of production. Price (personal communication, 2003) reports that the Shuar collective action in sugar cane cultivation ultimately failed. Everyone who participated was guaranteed an equal share of the proceeds from selling the crop, and there were consensually agreed upon fines for not showing up to clear the fields. But the fines had no bite: instead of being levied after each work episode (each episode in which participation occurred and could be monitored), the fines were to be deducted from each individual's profit once the crop was harvested and sold. The iterative ratchet effect ensued. Over time, participation in the cultivation effort dwindled to the point where the project failed and there were no proceeds to share. It is worth

noting that everyday life among the Shuar involves norms promoting generosity and sharing at levels rarely seen in the West.

Communitarian methods of organizing production have a strong ethical pull for many people, including moral philosophers. Equal division of profits can seem fair (under the assumption that everyone is contributing equally) or at least humane (under the assumption that everyone who is capable of contributing is doing so). The fairness of these compensation schemes is predicated on the assumption that no one free rides. Their efficacy is predicated on the assumption that if free riding does occur, contributors will continue to work at the same level—there will be no iterative ratchet effect. Are these reasonable assumptions? Ethicists need to consider whether certain methods of compensation invite free riding and dwindling participation, given the kind of minds we have.

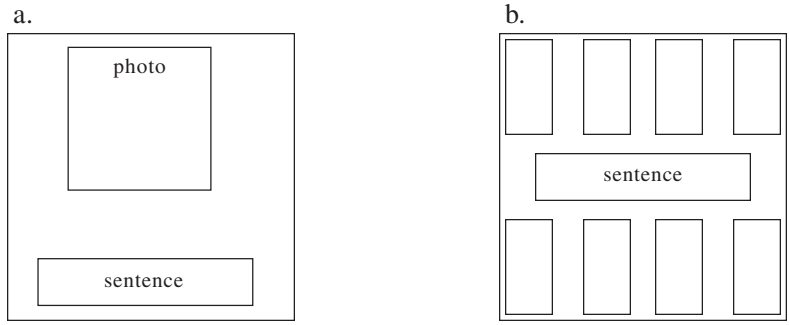
Farms, factories, restaurants—all involve multi-individual cooperation and hence collective action. The question is, are these projects organized as public goods (everyone benefits equally, regardless of their level of participation)? Or are payoffs organized such that effort is rewarded and free riding is punished? In the former Soviet Union, three percent of the land on collective farms was held privately, so local farming families could grow food for their own consumption and privately sell any excess. Yet estimates at the time were that this three percent of land produced forty-five percent to seventy-five percent of all the vegetables, meat, milk, eggs, and potatoes consumed in the Soviet Union (Sakoff, 1962). The quality of land on the collectively-held plots was the same; their low productivity was due to the iterative ratchet effect. People shifted their efforts away from the collective to the private plots. Without these private plots, it is likely that the people of the Soviet Union would have starved. Would this outcome have been ethically acceptable? Is a compensation procedure humane if its predictable consequence is mass suffering?

Workplace Diversity: When (*and* Why) Do *People* Notice *and* Remember Race?

Any given individual is a member of many different social categories: Leslie might be a boss, an engineer, a woman, a wife, a mother, an African-American, a church-goer. But when you meet Leslie, what do you notice and remember about her? That is, which category memberships do you encode?

Social psychologists can tell, using a memory confusion protocol (Taylor, Fiske, Etcoff, and Ruderman, 1978). This method uses errors in recall to unobtrusively reveal whether subjects are categorizing target individuals into groups and, if so, what dimensions they are using to do so (see Figure 2, p. 116). This method has revealed that, when adults encounter a new individual, they encode that individual's race, sex, and age (Taylor et al., 1978; Hewstone, Hantzi, and Johnston, 1991; Stangor, Lynch, Duan, and Glass, 1992; for review and discussion, see Brewer, 1988; Fiske and Neuberg, 1990; Hamilton, Stroessner, and Driscoll, 1994; Messick and Mackie, 1989). These dimensions can be encoded

Figure 2



Legend for Figure 2. The memory confusion protocol uses errors in recall to unobtrusively reveal whether subjects are categorizing target individuals into groups and, if so, what dimensions they are using to do so. Subjects are asked to form impressions of individuals whom they will see engaged in a conversation. They then see a sequence of sentences, each of which is paired with a photo of the individual who said it (see panel a). Afterwards, there is a surprise recall task: the sentences appear in random order, and subjects must attribute each to the correct individual (see panel b). Misattributions reveal encoding: Subjects more readily confuse individuals whom they have encoded as members of the *same* category than those whom they have categorized as members of *different* categories. For example, a citizen of Verona who had encoded coalition membership would make more within-category errors—errors in which s/he confused, say, a Capulet with a Capulet (or a Montague with a Montague)—than between-category errors—ones in which s/he confused a Capulet with a Montague or vice versa (this relationship will hold for data that are corrected to equalize base rates).

without other individuating information; for example, one might recall that one's new client or colleague is a young, white woman, without remembering anything else about her—her name, her hair color, her hometown.

Until recently, it appeared that race—along with sex and age—was encoded in an automatic and mandatory fashion. The encoding of race was thought to be spontaneous and automatic because the pattern of recall errors that indicates race encoding occurred in the absence of instructions to attend to the race of targets, and across a wide variety of experimental situations. It was thought to be mandatory—encoded with equal strength across all situations—because every attempt to increase or decrease the extent to which subjects encode the race of targets had failed (Taylor et al., 1978; Hewstone et al., 1991; Stangor et al., 1992). Until recently, no context manipulation—whether social, instructional, or attentional—had been able to budge this race effect. Such results led some to propose that race (along with sex and age) is a "primary" or "primitive" dimension of person perception, built into our cognitive architecture (e.g., Messick and Mackie, 1989; Hamilton et al., 1994).

Automatic Race Encoding is a Puzzle

For millions of years, our ancestors inhabited a social world in which registering the sex and life-history stage of an individual would have enabled a large variety of useful probabilistic inferences about that individual (e.g., adolescent girl; toddler boy). So natural selection could have favored neurocomputational machinery that automatically encodes an individual's sex and age. But "race" is a different matter.

Ancestral hunter-gatherers traveled primarily by foot, making social contact geographically local (Kelly, 1995). Given the breeding structure inherent in such a world, the typical individual would almost never have encountered people drawn from populations genetically distant enough to qualify as belonging to a different "race" (even if one could make biological sense of the concept; geneticists have failed to discover objective patterns in the world that could easily explain the racial categories that seem so perceptually obvious to adults; for reviews, see Hirschfeld, 1996; Cosmides, Tooby, and Kurzban, 2003). If individuals typically would not have encountered individuals of other races, then there could have been no selection for cognitive adaptations designed to preferentially encode such a dimension, much less encode it in an automatic and mandatory fashion.

For this reason, "race" is a very implausible candidate for a conceptual primitive to have been built into our evolved cognitive machinery. Race encoding may be a robust and reliable phenomenon, but it cannot be caused by computational machinery that was designed by natural selection for that purpose. This means that race encoding must be a side-effect of machinery that was designed by selection for some alternative function. If that machinery and its function are known, one might be able to create social contexts that diminish or eliminate race encoding.

Encoding Coalitional Alliances

In our view, no part of the human cognitive architecture is designed specifically to encode race. With our colleague, Robert Kurzban, we hypothesized that encoding of race is a byproduct of adaptations that evolved for an alternative function that was a regular part of the lives of our foraging ancestors: detecting coalitions and alliances (Kurzban, Tooby, and Cosmides, 2001). Hunter-gatherers lived in bands, and neighboring bands frequently came into conflict with one another (Ember, 1978; Manson and Wrangham, 1991; Keeley, 1996). Similarly, there were coalitions and alliances within bands (Chagnon, 1992), a pattern found in related primate species and likely to be far more ancient than the hominid lineage itself (Smuts et al. 1987; Wrangham and Peterson, 1996). To negotiate their social world successfully, anticipating the likely social consequences of alternative courses of action, our ancestors would have benefited by being equipped with neurocognitive machinery that tracked these shifting alliances.

Tracking alliances. Consider a program designed to infer who is allied with whom under ancestral conditions. What clues might that program use to do this? What factors in the world should it encode?

Alliance tracking mechanisms should notice patterns of coordinated action, cooperation, and competition. This is the primary database from which alliances can be inferred. But acts of cooperation and competition—behaviors that reveal one's coalitional allegiances—do not occur all the time. Like all behaviors, they are transitory. Alliance tracking machinery could form a better map of the political landscape if it were designed to use these rare revelatory behaviors to isolate additional cues that are correlated with coalitional behavior, but are more continuously present and perceptually easier to assay. This cue-mapping would allow one to use the behavior of some people to predict what others are likely to do.

Cues come in many forms. Some are intentional markers of one's coalitional alliances: war paint, gang colors, political buttons, for example. Other cues are incidental markers. Ethnographically well-known examples include accent and dialect, manner, gait, customary dress, family resemblance, and ethnic badges. If alliance tracking programs detect correlations between allegiance and appearance, then stable dimensions of shared appearance—which may be otherwise meaningless—would emerge in the cognitive system as markers of social categories. Coalitional computation would increase their subsequent perceptual salience, and encode them at higher rates. Any readily observable feature—however arbitrary—should be able to acquire social significance and cognitive efficacy when it validly cues patterns of alliance.

Modern conditions. In societies that are not completely racially integrated, shared appearance—a highly visible and always present cue—can be correlated with patterns of association, cooperation, and competition (Sidanius and Pratto, 1999). Under these conditions, coalition detectors may perceive (or misperceive) race-based social alliances, and the mind will map "race" onto the cognitive variable *coalition*. According to this hypothesis, race encoding is not automatic and mandatory. It appeared that way only because the relevant research was conducted in certain social environments where the construct of "race" happened, for historical reasons (Hirschfeld, 1996), to be one valid probabilistic cue to a different underlying variable, one that the mind *was* designed to automatically seek out: coalitional affiliation (Kurzban, 2001; Kurzban, Tooby, and Cosmides, 2001; Sidanius and Pratto, 1999; Tooby and Cosmides, 1988).

Dynamic revision. Patterns of alliance often change when new issues arise whose possible resolutions differentially affect new subsets of the local social world. Consequently, coalitions shift over time, varying in composition, surface cues, duration and internal cohesion. To track these changes, cue validities would need to be computed and revised dynamically: No single coalitional cue (including cues to race) should be uniformly encoded across all contexts. Furthermore, arbitrary cues (such as skin color) should pick up—and lose—significance only insofar as they acquire predictive validity for coalitional membership.

There is a direct empirical implication of the hypothesis that race is encoded by alliance tracking machinery and that this machinery dynamically updates coalition cues to keep up with new situations. Consider a salient coalitional conflict in which race is *not* correlated with coalition membership. Two things should happen: (i) Arbitrary shared appearance cues that do predict coalition membership should be strongly encoded, and (ii) race encoding should decrease.

Is *Coalition* Encoded?

Using the memory confusion protocol, Kurzban, Tooby, and Cosmides (2001) confirmed both predictions. They first showed that people do automatically encode the coalitional alliances of targets. The targets were males, some black, some white; each made statements suggesting allegiance with one of two antagonistic coalitions. Crucially, race was not correlated with coalitional affiliation.

Subjects encoded coalitional alliance even in the absence of shared appearance cues—merely from patterns of agreement and disagreement. But when a shared appearance cue—jersey color—was added, coalition encoding was boosted dramatically, to levels higher than any found for race. (N.B. Jersey color is not encoded at all when it lacks social meaning [Stangor et al., 1992].)

Race as a Proxy for Coalition?

The results further showed that, as predicted, race encoding is not mandatory. When coalition encoding was boosted by a shared appearance cue, there was an accompanying decrease in race encoding, which was diminished in one experiment and eliminated in another. Other tests showed that the decrease in race encoding could not be attributed to domain-general constraints on attention.

Subjects had a lifetime's experience of race predicting patterns of cooperation and conflict. The decreases in these experiments occurred in response to only 4 minutes of exposure to an alternative world where race did not predict coalitional alliance. This is expected if (i) race is encoded (in real life) because it serves as a rough-and-ready coalition cue, and (ii) coalition cues are revised dynamically, to reflect newly emerging coalitions. There are many contexts that decrease racial *stereotyping* (inferences); creating alliances uncorrelated with race is the first social context found that decreases race *encoding*.

The results suggest that the tendency to notice and remember "race" is an easily reversible byproduct of programs that detect coalitional alliances. When the relevant coalitional conflict was uncorrelated with race, the tendency to notice and remember the race of the individuals involved diminished, and sometimes even disappeared. These results suggest that the social construct "race" is a byproduct of programs that evolved to look not for race *per se*, but for coalitional alliances. In a sense, the creation of multiracial coalitions "erased race" in the minds of our subjects.

Implications for Business Ethics: Harmony in the Workplace

There is no doubt that racial discrimination occurs; for an alarming and fascinating compendium of data on this point, we recommend *Social Dominance*, by Sidanius and Pratto (1999; their theoretical analysis is both interesting and relevant to business ethicists). The question is, how can the situation be improved?

It is often assumed that the way to promote harmonious cooperation in the workplace is to first eradicate racial stereotypes. Sensitivity training courses are created, the goal of which is to make people of one race aware of the negative inferences they make about people of another race, and to highlight the overt and subtle ways in which individuals of that race are discriminated against. Note, however, that this very process divides people into two different camps (e.g., "whites" and "blacks"), a process that social psychologists know promotes ingroup favoritism and outgroup derogation (Sherif et al. 1961; Tajfel et al. 1971; Brewer, 1979). It *reinforces* the social construction of racial groups as opposing coalitions. However well-intentioned sensitivity courses may be, if race is a proxy for coalition, they may turn out to exacerbate racial problems rather than mitigate them.

Kurzban, Tooby, and Cosmides' results suggest an intriguing alternative. Instead of trying to eradicate racism to get cooperation, companies might be able to use cooperation to eradicate racism. In making and marketing a product, companies create many small task-oriented teams: multi-individual cooperative coalitions. Creating teams where race does not predict team membership should decrease attention to race. At least that is what happened in the Kurzban experiments: when coalition membership could not be predicted by race, there was a decrease in people's attention to race.

We do not yet know the parameters and boundaries of Kurzban et al.'s "erasing race" effect. For example, do the coalitions have to be perceived as in conflict with one another, or does each merely have to be composed of individuals coordinating their behavior with one another in pursuit of a common goal? The creation of multiracial corporate teams could have the desired effect, however, even if it turns out that an element of conflict is necessary—that is, even if erasing race requires the construction of a (nonracial) *us* and *them*. After all, the fact that every company has competitors in the marketplace provides the raw material for thinking in terms of opposing "teams," as surely as the NFL does: Our marketing team versus theirs, our sales department versus theirs. The creation of groups where race does not predict alliances may be a way that companies can improve race relations and diversity, a method that works with human nature rather than against it.

Conclusions

What is ethically achievable depends on what is humanly possible. But what *is* humanly possible? When people think about limits to human action, external constraints come to mind: a stone too heavy to lift, a bullet too fast to dodge. But human action—behavior—is muscle movements, and muscle movement is caused by sophisticated cognitive programs. If this is unclear, consider that most movement disorders—paralysis, epileptic seizures, the tics and shouted epithets of Gilles de Tourette's syndrome, the shakes of Parkinson's disease—are caused by injury to or disorders of the cognitive programs that move muscles, not injury to the muscles themselves (Frith, 1992). In discussing what actions are humanly possible, we should start taking the cognitive programs that *cause* action into account.

Some of these are motivational programs, designed to start and stop our muscles in particular ways as a function of the situations we face. Standing still while watching a lion lunge toward one's throat may be impossible for a normal, brain intact human being. Our motivational systems were *designed* to mobilize evasive action in response to a lunging lion, and no sane ethical system would exhort a person to do otherwise.

Similarly, the human mind may have motivational systems that are *designed* to lower the amount of effort one expends on a collective action as a function of whether others are free riding. There may be no way to over-ride this gumption drain except by applying extreme coercive force—and thus engaging a different motivational system. Yet threatening someone's life is not an ethically neutral course of action. A better solution might be to avoid organizing the project as a public good in the first place.

The human mind may have a computational system that is *designed* to construct an *us* and a *them* under certain circumstances. Moral exhortation may not be enough to counteract the consequences of this social construction. Indeed, it may drive negative attitudes toward *them* underground, changing what people profess without changing their attitudes (Greenwald and Banaji, 1995). A better solution might be to create a new "us": a team comprised of Capulets *and* Montagus, of "blacks" *and* "whites." The experience of participating in such a coalition may diminish attention to the old divisions and reshape attitudes spontaneously, without moral exhortation.

The human mind may be attuned to detecting cheaters yet lack systems designed to seek out other kinds of ethical violations. Penalizing failures to notice such violations may be useless. A better solution might be to reframe ethical rules to take advantage of the human mind's natural abilities to detect cheaters and to attend to violations of precautionary rules.

Soft versus hard solutions. In their attempts to create a more just and ethical society, most people take the approach advocated by Katherine Hepburn's character, Rose Sayer, in *The African Queen*. When Charlie Allnut (Humphrey Bogart) tries to excuse some piece of bad behavior by saying "it's only human

nature," Hepburn replies "Nature, Mr. Allnut, is what we are put in this world to rise above." Her words conjure the image of a Manichean struggle, in which "willpower" is deployed to counteract the low and degrading forces of our evolved human nature.

We think this is a losing strategy. Our minds are indeed equipped with over-ride programs—willpower, if you will (Baron-Cohen, Robertson, and Moriarty, 1994; Frith, 1992). But if override is necessary, the battle is already half lost. Far better are "soft" policies, solutions like the ones suggested above. These create situations that activate and deactivate the evolved programs that motivate ethical and unethical behavior.

Why call these solutions "soft"? Boxers are trained to meet the opponent's moves with countervailing force: a "hard," Hepburn approach. In soft martial arts, like aikido, one is trained to achieve goals by exploiting the moves that the opponent is already making. Equipped with an understanding of human nature, it may be possible to train people in ethical aikido: the art of designing policies that achieve ethical goals by taking advantage of the moves that our human nature is already prepared to make. But to do this, we must first know ourselves.

Notes

This paper was first delivered in April 2002 as a Ruffin Lecture on Business Ethics and Science, at the Olsson Center for Applied Ethics, Darden School of Business Administration, University of Virginia. We thank Bill Frederick, Ed Freeman, and the participants for many stimulating conversations. We also thank Douglass North and the Mercatus Center for inviting us to participate in their Social Change workshops; the ideas and excitement of those workshops informed many of our thoughts herein. We are especially grateful to Vernon Smith, who introduced us to the field of experimental economics and showed us its relevance to research in evolutionary psychology.

1. Ethology integrated advances #2 and #3; sociobiology integrated 2-4; evolutionary psychology integrates 1-4 into the framework described in C-1 through C-6.

2. Most choose either *P* alone, or *P & Q*. They are not reasoning correctly from a biconditional interpretation; if they were, they would choose all four cards (a rare response).

3. And when the rule does not activate an alternative cognitive adaptation, as precautionary rules do; see Fiddick, 1998; Fiddick, Cosmides & Tooby, 2000).

References

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

Axelrod, R., and W. D. Hamilton. 1981. The evolution of cooperation. *Science* 211: 1390–1396.

Baron-Cohen, S. 1995. *Mindblindness: An essay on autism and theory of mind.* Cambridge, Mass.: MIT Press.

Baron-Cohen, S., M. Robertson, and J. Moriarty. 1994a. The development of the will: A neuropsychological analysis of Gilles de la Tourette's Syndrome. In *The self and its dysfunction: Proceedings of the 4th Rochester symposium*, ed. D. Cicchetti and S. Toth. Rochester, N.Y.: University of Rochester Press.

Barrett, H. C. 1999. Guilty minds: How perceived intent, incentive, and ability to cheat influence social contract reasoning. 11th Annual Meeting of the *Human Behavior and Evolution Society*, Salt Lake City, Utah.

Bentham, J. 1789. *An introduction to the principles of morals and legislation*. London: T. Payne.

Boyd, R. 1988. Is the repeated prisoners' dilemma a good model of reciprocal altruism? *Ethology and Sociobiology* 9: 211–222.

Brewer, M. 1979. Ingroup bias in the minimal intergroup situation: A cognitive motivational analysis. *Psychological Bulletin* 86: 307–324.

_____. 1988. A dual process model of impression formation. In *Advances in Social Cognition* 1, ed. T. Srull and R. Wyer: 1–36.

Buss, D. M. 1989. Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences* 12: 1–49.

Cashdan, E. 1989. Hunters and gatherers: Economic behavior in bands. In *Economic Anthropology*, ed. S. Plattner. Stanford, Calif.: Stanford University Press.

Chagnon, N. 1992. *Yanomamo (4th edition).* Fort Worth, Tex.: Harcourt.

Cheng, P., and K. Holyoak. 1985. Pragmatic reasoning schemas. *Cognitive Psychology* 17: 391–416.

Cosmides, L. 1985. Deduction or Darwinian Algorithms? An explanation of the "elusive" content effect on the Wason selection task. Doctoral dissertation, Harvard University. *University Microfilms #86-02206.*

_____. 1989. The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* 31: 187–276.

Cosmides, L., and J. Tooby. 1989. Evolutionary psychology and the generation of culture, Part II. Case study: A computational theory of social exchange. *Ethology & Sociobiology* 10: 51–97.

_____. 1992. Cognitive adaptations for social exchange. In *The adapted mind: Evolutionary psychology and the generation of culture*, ed. J. Barkow, L. Cosmides, and J. Tooby. New York: Oxford University Press.

_____. 1997. Dissecting the computational architecture of social inference mechanisms. In *Characterizing human psychological adaptations* (Ciba Foundation Symposium #208). Chichester: Wiley: 132–156.

_____. 2000a. Consider the source: The evolution of adaptations for decoupling and metarepresentation. In *Metarepresentations: A multidisciplinary perspective*, ed. D. Sperber. Vancouver Studies in Cognitive Science. New York: Oxford University Press: 53–115.

_____. 2000b. The cognitive neuroscience of social reasoning. In *The New Cognitive Neurosciences, Second Edition* (chapter 87), ed. M. S. Gazzaniga. Cambridge, Mass.: MIT Press: 1259–1270.

Cosmides, L., J. Tooby, and R. Kurzban. In press, 2003. Perceptions of race. *Trends in Cognitive Sciences*.

Daly, M., and M. Wilson. 1988. *Homicide*. New York: Aldine.
Ember, C. R. 1978. Myths about hunter-gatherers. *Ethnology* 27: 239–448.
Fehr, E., and S. Gächter. 2000a. Cooperation and punishment in public goods experiments. *American Economic Review* 90: 980–994.
_____. 2000b. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* 14: 159–181.
Fiddick, L. 1998. *The deal and the danger: An evolutionary analysis of deontic reasoning*. Doctoral dissertation, Department of Psychology, University of California, Santa Barbara.
Fiddick, L. 2004 (in press). Domains of deontic reasoning: Resolving the discrepancy between the cognitive and moral reasoning literatures. *The Quarterly Journal of Experimental Psychology*, 57A.
Fiddick, L., L. Cosmides, and J. Tooby. 2000. No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition* 77: 1–79.
Fiske, S., and S. Neuberg. 1990. A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in Experimental Social Psychology* 23, ed. M. Zanna (Academic Press): 1–74.
Frith, C. 1992. *The Cognitive Neuropsychology of Schizophrenia.* Hillsdale, N.J.: Erlbaum.
Gigerenzer, G. 2000. *Adaptive thinking: Rationality in the real world*. New York: Oxford.
Greenwald, A. G., and M. R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102: 4–27.
Gurven, M. In press. To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences* (http://www.bbsonline.org/Preprints/Gurven-06282002/Gurven.pdf).
Gurven, M., K. Hill, H. Kaplan, A. Hurtado, and R. Lyles. 2000. Food transfers among Hiwi foragers of Venezuela: Tests of reciprocity. *Human Ecology* 28(2): 171–218.
Hamilton, D., S. Stroessner, and D. Driscoll. 1994. Social cognition and the study of stereotyping. In *Social cognition: Impact on social psychology*, ed. P. G. Devine, D. Hamilton, and T. Ostrom. San Diego: Academic Press: 291–321.
Hamilton, W. D. 1964. The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7: 1–52.
Hewstone, M., A. Hantzi, and L. Johnston. 1991. Social categorization and person memory: The pervasiveness of race as an organizing principle. *European Journal of Social Psychology* 21: 517–528.
Hirschfeld, L. 1996. *Race in the Making*. Cambridge, Mass.: MIT Press.
Isaac, G. 1978. The food-sharing behavior of protohuman hominids. *Scientific American* 238: 90–108.
Keeley, L. 1996. *War before civilization: The myth of the peaceful savage*. Oxford: Oxford University Press.
Kelly, R. 1995. *The Foraging Spectrum*. Washington, D.C.: Smithsonian Institution Press.

Kurzban, R. 2001. The social psychophysics of cooperation: Nonverbal communication in a public goods game. *Journal of Nonverbal Behavior* 25: 241–259.

Kurzban, R., K. McCabe, V. Smith, and B. J. Wilson. 2001. Incremental commitment and reciprocity in a real time public goods game. *Personality and Social Psychology Bulletin* 27(12): 1662–1673.

Kurzban, R., J. Tooby, and L. Cosmides. 2001. Can race be erased?: Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences* 98(26) (December 18, 2001): 15387–15392.

Leckman, J., and L. Mayes. 1998. Maladies of love—an evolutionary perspective on some forms of obsessive-compulsive disorder. In *Advancing research on developmental plasticity: Integrating the behavioral science and neuroscience of mental health*, ed. D. M. Hann, L. C. Huffman, I. I. Lederhendler, and D. Meinecke. Rockville, Md.: NIMH, US Dept. of Health and Human Services: 134–152.

_____. 1999. Preoccupations and behaviors associated with romantic and parental love: Perspectives on the origin of obsessive-compulsive disorder. Obsessive-Compulsive Disorder 8(3): 635–665.

Lee, R., and I. DeVore, eds. 1968. *Man the Hunter*. Chicago: Aldine.

Leslie, A. 1987. Pretense and representation: The origins of "theory of mind." *Psychological Review* 94, 412–426.

Manktelow, K., and D. Over. 1991. Social roles and utilities in reasoning with deontic conditionals. *Cognition* 39: 85–105.

Manson, J., and R. Wrangham. 1991. Intergroup aggression in chimpanzees and humans. *Current Anthropology* 32(4): 369–390.

Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.

Messick, D. and D. Mackie. 1989. Intergroup relations. *Annual Review of Psychology* 40: 45–81.

North, D. 1990. *Institutions, Institutional Change and Economic Performance*. New York: Cambridge University Press.

Nozick, R. 1975. *Anarchy, State, and Utopia*. Cambridge, Mass.: Harvard University Press.

Olson, M. 1965. *Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Mass.: Harvard University Press.

Packer, C. 1977. Reciprocal altruism in *Papio annubis*. *Nature* 265: 441–443.

Pinker, S. 1994. *The language instinct*. New York: HarperCollins.

_____. 1997. *How the mind works*. New York: Norton.

_____. 2002. *The Blank Slate*. New York: Norton.

Platt, R., and R. Griggs. 1993. Darwinian algorithms and the Wason selection task: A factorial analysis of social contract selection task problems. *Cognition* 48: 163–192.

Price, M. E., L. Cosmides, and J. Tooby. 2002. Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior* 23: 203–231.

Price, M., H. C. Barrett, and E. Hagen. Under review. Collective action and punishment in the Ecuadorian Amazon.

Rawls, J. 1971. *A theory of justice*. Cambridge, Mass.: Harvard University Press.

Sakoff, A. 1962. The private sector in Soviet agriculture. *Monthly Bulletin of Agricultural Economics* 11: 9.

Sen, A. 1999. *Development as freedom*. New York: Knopf.

Sherif, M., O. Harvey, B. White, W. Hood, and C. Sherif. 1961. *Intergroup conflict and cooperation: The robbers cave experiment*. Norman: University of Oklahoma Book Exchange.

Sidanius, J., and F. Pratto. 1999. *Social Dominance*. New York: Cambridge University Press.

Smuts, B., D. Cheney, R. Seyfarth, R. Wrangham, and T. Struhsaker. 1987. *Primate Societies*. Chicago: University of Chicago Press.

Sperber, D., F. Cara, and V. Girotto. 1995. Relevance theory explains the selection task. *Cognition* 57, 31–95.

Stangor, C., L. Lynch, C. Duan, and B. Glas. 1992. Categorization of individuals on the basis of multiple social features. *Journal of Personality & Social Psychology* 62: 207–218.

Stone, V., L. Cosmides, J. Tooby, N. Kroll, and R. Knight, R. (2002). Selective Impairment of Reasoning About Social Exchange in a Patient with Bilateral Limbic System Damage. *Proceedings of the National Academy of Sciences* (August, 2002).

Sugiyama, L., J. Tooby, and L. Cosmides. 2002. Cross-cultural evidence of cognitive adaptations for social exchange among the Shiwiar of Ecuadorian Amazonia. *Proceedings of the National Academy of Sciences* (August, 2002).

Tajfel, H., M. Billig, R. Bundy, and C. Flament. 1971. Social categorization and intergroup behavior. *European Journal of Social Psychology* 1: 149–178.

Taylor, S., S. Fiske, N. Etcoff, and A. Ruderman. 1978. Categorical bases of person memory and stereotyping. *Journal of Personality and Social Psychology.* 36: 778–793.

Tooby, J., and L. Cosmides. 1988. The evolution of war and its cognitive foundations. *Institute for Evolutionary Studies Technical Report* 88-1.

_____. 1990. On the universality of human nature and the uniqueness of the individual: The role of genetics and adaptation. *Journal of Personality* 58: 17–67.

_____. 1996. Friendship and the Banker's Paradox: Other pathways to the evolution of adaptations for altruism. In *Evolution of Social Behaviour Patterns in Primates and Man. Proceedings of the British Academy* 88, ed. W. G. Runciman, J. Maynard Smith, and R. I. M. Dunbar: 119–143.

Trivers, R. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46: 35–57.

de Waal, F. 1989. Food sharing and reciprocal obligations among chimpanzees. *Journal of Human Evolution* 18: 433–459.

Wason, P. 1966. Reasoning. In *New Horizons in Psychology*, ed. B. Foss. Harmondsworth: Penguin.

_____. 1983. Reasoning and Rationality in the selection task. In *Thinking and Reasoning: Psychological Approaches,* ed. J. S. B. T. Evans. London: Routledge & Kegan Paul.

Wason, P., and P. Johnson-Laird. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, Mass.: Harvard University Press.

Wilkinson, G. 1988. Reciprocal altruism in bats and other mammals. *Ethology and Sociobiology* 9: 85–100.

Wrangham, R., and D. Peterson. 1996. *Demonic males: Apes and the origins of human violence*. Boston: Houghton Mifflin.

Yamagishi, T. 1986. The provision of a sanctioning system as a public good. *Journal of Personality & Social Psychology* 51, 110–116.