# Evolutionary Psychology and the Generation of Culture, Part II

## Case Study: A Computational Theory of Social Exchange

Leda Cosmides and John Tooby
*Department of Psychology, Stanford University, Stanford, California*

**Models of the various adaptive specializations that have evolved in the human psyche could become the building blocks of a scientific theory of culture. The first step in creating such models is the derivation of a so-called "computational theory" of the adaptive problem each psychological specialization has evolved to solve. In Part II, as a case study, a sketch of a computational theory of social exchange (cooperation for mutual benefit) is developed. The dynamics of natural selection in Pleistocene ecological conditions define adaptive information processing problems that humans must be able to solve in order to participate in social exchange: individual recognition, memory for one's history of interaction, value communication, value modeling, and a shared grammar of social contracts that specifies representational structure and inferential procedures. The nature of these adaptive information processing problems places constraints on the class of cognitive programs capable of solving them; this allows one to make empirical predictions about how the cognitive processes involved in attention, communication, memory, learning, and reasoning are mobilized in situations of social exchange. Once the cognitive programs specialized for regulating social exchange are mapped, the variation and invariances in social exchange within and between cultures can be meaningfully discussed.**

KEY WORDS: Reciprocal Altruism; Cooperation; Tit for tat; Cognition; Reasoning; Evolution; Learning; Culture.

## INTRODUCTION

Human beings live in groups, and their behavior is affected by information derived from the other individuals with whom they live. The study of culture is the study of how different kinds of information from each individual's environment, especially from

Received April 1, 1987; revised October 17, 1987.

Address reprint requests to: Leda Cosmides, Department of Psychology, Bldg. 420, Stanford University, Stanford, CA 94305.

Ethology and Sociobiology 10: 51–97 (1989)

655 Avenue of the Americas, New York, NY 10010
0162-3095/89/$3.50

his or her social environment, can be expected to affect that individual's behavior. The behavior elicited by this information reverberates throughout the individual's social group, as information that other individuals may act on in turn. The ongoing cycle that results is the generation of culture. By directly regulating individual learning and behavior, those psychological mechanisms that select and process information from the individual's social environment govern the resulting cultural dynamics. The key to understanding cultural processes therefore lies in the discovery, and subsequent mapping, of the properties of these innate information processing mechanisms, an enterprise now feasible using an approach that is both evolutionary and cognitive (Tooby and Cosmides this issue).

Evolutionary biology provides a powerful heuristic for guiding psychological research. This heuristic rests on the recognition that psychological mechanisms evolved as responses to selection pressures. The more important the adaptive problem, the more intensely selection will have specialized and improved the performance of these mechanisms. Some of these mechanisms evolved to meet the adaptive problem of social exchange. We will use the case of social exchange to illustrate the method of evolutionary psychology discussed in Part I of "Evolutionary Psychology and the Generation of Culture" (Tooby and Cosmides this issue).

Social exchange—cooperation between two or more individuals for mutual benefit—is a pervasive aspect of all human cultures. It manifests itself in many different forms within and across cultures: ritualized gift giving, the exchange of favors between friends, trade in a market economy, and barter are all forms of social exchange. Successfully conducted social exchange was a critically important feature of hominid evolution (Tooby and DeVore 1987). The ability to successfully engage in social exchange depends on the structured processing of specific kinds of information. This is because natural selection permits the evolution of only certain strategies for engaging in social exchange: Some strategies are evolutionarily stable (Maynard Smith 1982), others are not. By studying the nature of these strategies, one can deduce many properties that the human information processing algorithms regulating social exchange must have, as well as much about the associated capabilities such algorithms require to function properly. Using this framework, one can then make empirical predictions about human performance in areas that are the traditional concern of cognitive psychologists: attention, communication, the organization of memory, learning, and reasoning.

By examining of the nature of the selective pressures on social exchange in human evolution, some things can be inferred about the psychological basis for social exchange in humans. These inferences can be used to construct a computational theory of social exchange: a theory defining the information processing problems that a human engaging in social exchange must be able to solve. A psychological mechanism is a solution to an information processing problem. Thus, a computational theory provides a pre-

dictive framework that facilitates the design of experiments that can map the structure of the cognitive programs that guide social exchange in humans.

This article is divided into four sections:

Part 1. Only certain strategies for engaging in social exchange can evolve: Natural selection's game theoretic structure defines what properties these strategies must have.

Part 2. The ecological conditions necessary for the evolution of social exchange were manifest during hominid evolution; hominid behavioral ecology further constrains a computational theory of social exchange.

Part 3. These strategic and ecological constraints define a set of information processing problems that must be solved by any human engaging in social exchange. Computational theories of these problems are developed.

Part 4. Aspects of the computational theory of social exchange have been tested. Experimental evidence from tests of logical reasoning verifies the existence of algorithms for detecting cheaters; as predicted, these algorithms operate on item-independent, cost-benefit representations of exchange interactions.

Parts 1 and 2 review the constraints from which a computational theory should be built. Part 3 presents a first attempt to build a computational theory of social exchange; Part 4 briefly reviews the results of experiments designed to test aspects of the computational theory developed.

## PART 1. NATURAL SELECTION AND SOCIAL EXCHANGE

> The critical act in formulating computational theories turns out to be the discovery of valid constraints on the way the world is structured. . . . (Marr and Nishihara 1978)

There are laws inherent in the dynamics of natural selection that should shape every species. Many of these laws govern the evolution of social behavior; they constrain the kinds of social behavior that can evolve under a given set of circumstances. Although this is common knowledge among evolutionary biologists, it is something psychologists need to incorporate into their methodology: knowledge of natural selection constitutes knowledge of a set of valid constraints on what kinds of mental mechanisms could have evolved and on the properties they have.

Traits can be thought of as the embodiment of *strategies* for the propagation of the genes that code for them. By analyzing the dynamics of gene flow through populations, one can determine what kinds of traits will be quickly selected out and what kinds of traits are likely to become universal and species-typical. Formally, this analysis can be cast in terms of game theory: One strategy is pitted against others to see which ones come to

dominate the gene pool (Maynard Smith 1982). During the last 20 years, game-theoretic models of the dynamics of natural selection have proliferated in evolutionary biology. The elaboration of these methods now allows a more precise characterization of the differences between those strategies that can be selected for and those that will be selected against (e.g., Hamilton 1964; Williams 1966; Maynard Smith 1982; Dawkins 1982). In the case of social exchange ("cooperation" or "reciprocation"), Axelrod and Hamilton (1981) and Axelrod (1984) have shown that only certain families of strategies with certain distinctive properties can evolve.

Using the iterated Prisoner's Dilemma as their paradigm of cooperation,[1] Axelrod and Hamilton (1981), and Axelrod (1984), following Williams (1966) and Trivers (1971), explored the envelope of conditions limiting the evolution of social exchange. It is the possibility of cheating (or defecting once one side of a mutually beneficial exchange has been carried out) that makes the evolution of cooperation difficult. For this reason, indiscriminate cooperation under conditions that allow cheating is an unstable strategy that would be quickly selected out under all models of biologically plausible conditions. Virtually any nonsimultaneous exchange creates a possibility of defection, and most "natural" opportunities for exchange are not simultaneous. For example:

1. A common "item" of exchange between primates is protection from conspecifics or predators. Two or more individuals develop coalitional relationships for mutual defense, aggression, or protection (e.g., baboons: Hall and DeVore 1965; chimpanzees: Wrangham 1986; de Waal 1982). If one individual is attacked, and another comes to its defense, there is nothing the aided individual can do at that time to repay his rescuer. Reciprocation is possible only at another time, when the rescuer is himself attacked.
2. Interactants are foraging for patchy resources. One individual finds, for example, a patch containing more than can be easily eaten by itself, and gives a call to guide others to the patch (or returns and shares the resource with others). Repayment in kind, to be valued, would have to take place subsequently (e.g., chimpanzees: Goodall 1968, 1971; bats: McCracken and Bradbury 1981; Wilkinson 1984).
3. In hunter-gatherer meat sharing, kills may be larger than can be easily consumed by those who were directly in the hunt, and only irregularly obtained. The value of consuming the whole kill is less than the value of sharing the unneeded or less-needed portions with others, provided that the act is reciprocated at some future time (Lee and DeVore 1968). Again, repayment on the spot is both unlikely and not valuable.

[1] Other models of social exchange are possible, but they will not change the basic conclusion of this section: that reciprocation is necessary for the evolution of social exchange. For example, the Prisoner's Dilemma assumes that enforceable threats and enforceable contracts are impossibilities (Axelrod 1984), assumptions that are frequently violated in nature. The introduction of these factors would not obviate reciprocation—in fact, they would enforce it.

4. There is mounting evidence that a baboon male forms "special relationships" with a few lactating (and therefore infertile) females and their infants: He protects them from conspecifics and predators, and may subsequently obtain sexual access when the females wean their infants and become fertile again (e.g., Smuts 1982; Strum 1985). If this does constitute an exchange, the male's repayment, by necessity, comes at a much later time.

The opportunity for on-the-spot repayment or withdrawal of benefits in the face of cheating is rare in nature for several reasons:

1. The "items" of exchange are frequently acts that, once done, cannot be undone (e.g., protection from attack, alerting others to the presence of a food source).
2. Opportunities for simultaneous mutual aid are rare because the needs and abilities of organisms are rarely exactly and simultaneously complementary: The female baboon is not fertile when her infant needs protection, yet this is when the male's ability to protect is of most value to her.
3. On those occasions when repayment could be made simultaneously and in the same currency, declining marginal utilities makes the exchange senseless: If meat sharers both make a kill on the same day, neither benefits from the other's windfall.

Thus, in the absence of a widely accepted medium of exchange,[2] most exchanges do involve opportunities for defection.

A system of mutual cooperation cannot emerge in a one move Prisoner's Dilemma, because it is always in the interest of each player to defect (Luce and Raiffa 1957; see Fig. 1 for explanation). In fact, the argument is general to any known, fixed number of games (Luce and Raiffa 1957). However, selection pressures change radically when individuals play a *series* of Prisoner's Dilemma games. Mutual cooperation—and therefore social exchange—can emerge between two players when 1) there is a high probability that they will meet again, 2) neither knows for sure exactly how many times they will meet,[3] and 3) they do not value later payoffs by too much less than earlier payoffs (Axelrod and Hamilton 1981; Axelrod 1984). If the parties are making a series of moves rather than just one, one party's behavior on one move can influence the other's behavior on future moves. (For ease of explication, in the discussions of social exchange throughout this paper, the

[2] Indeed, such factors are exactly why it is so useful to have a medium of exchange. One party doesn't have to be able to provide the particular goods or services the other party wants because money can be converted into anything others are willing to exchange for it. Furthermore, money permits a simultaneous exchange, in which either party can, in fact, withhold the money if he or she suspects that the other is attempting to cheat.

[3] The game "unravels" if they do. If we both know we are playing three games, then we both know we will mutually defect on the last game. In practice, then, our second game is our last game. But we know that we will, therefore, mutually defect on that game, so, in practice, we are playing only one game. The argument is general to any known, fixed number of games (Luce and Raiffa 1957).

| | | you: *cooperate* | you: *defect* |
|---|---|---|---|
| me | *cooperate* | me: $R$ = B(me) - C(me)<br>you: $R$ = B(you) - C(you) | me: $S$ = C(me)<br>you: $T$ = B(you) |
| | *defect* | me: $T$ = B(me)<br>you: $S$ = C(you) | me: $P$ = 0(me)<br>you: $P$ = 0(you) |

**FIGURE 1.** Social exchange sets up a Prisoner's Dilemma: Payoff Schedule.

$B(X)$ = Benefit to $X$; $C(X)$ = Cost to $X$; $0(x)$ = $X$'s inclusive fitness is unchanged. $R$ = Reward for mutual cooperation; $T$ = Temptation to defect; $S$ = Sucker's payoff; $P$ = Punishment for mutual defection. $T > R > P > S$; $R > (T + S)/2$ (for an iterated game; this prevents players from "cooperating" to maximize their utility by alternately defecting on one another. $C$(me) need not equal $C$(you), and $B$(me) need not equal $B$(you); an exchange will have the structure of a Prisoner's Dilemma as long as mutual cooperation would produce a net benefit for both players. Let us assume that you and I are playing a one move Prisoner's Dilemma game. I would reason thus: "You will either cooperate or defect. If you cooperate, then I get a higher payoff by defecting, because $T$, the Temptation to defect, is greater than $R$, the reward I would get for mutual cooperation. If you defect, then I get a higher payoff by also defecting, because $P$, the payoff I receive as a Punishment for mutual defection, is greater than $S$, the Sucker's payoff I would get if I cooperate and you defect. Therefore, no matter what you do, I am better off defecting." Your reasoning process would be identical, so we would both defect, and we would both get $P$, the Punishment for mutual defection. If you cooperate, I get $B$(me) for defecting instead of $B$(me) − $C$(me) for cooperating. If you defect, I lose nothing by defecting instead of losing $C$(me) by cooperating.

two interactants in the social exchange will be designated "you" and "I," with appropriate possessive pronouns). If "I" defect when "you" cooperated, then you can retaliate by defecting on the next move;[4] if I cooperate, you can reward me by cooperating on the next move. In an iterated Prisoner's Dilemma, a system can emerge that has incentives for cooperation and disincentives for defection.

The work of Trivers (1971), Axelrod and Hamilton (1981), and Axelrod (1984), has shown that indiscriminate cooperation cannot be selected for when the opportunity for cheating exists. A cooperative strategy can invade

[4] In nature, I can also retaliate by inflicting a cost on you through the use of violence. However, if I can, reliably, do this, the game is no longer a Prisoner's Dilemma. Violent retaliation is a "tax" on defection that wipes out the incentive to defect (i.e., $T$ minus $R$). If $T \leq R$, then the situation no longer presents a dilemma—we both have an incentive to cooperate and no incentive to cheat. The key word in the above scenario is *reliably*. From a "veil of ignorance" as to the relative strength of two individuals, on average, half the time I (the cheated on) will be able to inflict a cost on you, and half the time you (the cheater) will be able to inflict a cost on me. Of course, most animals are not acting from a veil of ignorance, and one would expect them to assess their relative strength and adjust their strategies accordingly.

a population of noncooperators if, and only if, it cooperates with other cooperators and excludes (or retaliates against) cheaters. If a cognitive decision rule regulating when one should cooperate and when one should cheat does not instantiate this constraint, then it will be selected against. However, Axelrod (1984) has shown that there are many decision rules that do instantiate this constraint. Any of these could (other things being equal), therefore, have been selected for in humans; which decision rule, out of this constrained family, actually evolved in the human lineage is an empirical question. The most general statement about such decision rules that natural selection theory permits is this: Humans have the ability to cooperate for mutual benefit; this capacity could not have evolved unless it included algorithms for detecting, and being provoked by, cheating.

## PART 2. SOCIAL EXCHANGE AND THE PLEISTOCENE ENVIRONMENT

> The ecological conditions necessary for the evolution of social exchange were manifest during hominid evolution; hominid behavioral ecology further constrains a computational theory of social exchange.

Cooperation can evolve only when 1) there are many situations in which individuals can benefit each other at relatively low cost to themselves (i.e., an *iterated* Prisoner's Dilemma game is possible), and 2) the probability of two individuals meeting again is sufficiently high.[5] The probability that two individuals will meet again is increased if the individuals are long-lived and have low dispersal rates. These life-history factors also increase the *number* of situations for mutual help that two individuals are likely to encounter. The ecological and life-history factors characteristic of the human environment of evolutionary adaptedness fulfill the conditions necessary for the evolution of cooperation. Pleistocene hunter-gatherers were not only long-lived, but they lived in small, relatively stable bands. Thus, the probability was high that an individual you had helped would be around when you needed help. Moreover, in all probability these individuals, like modern hunter-gatherers, were closely related; kin selection can promote the evolution of cooperation (Trivers 1971; Axelrod and Hamilton 1981).

The intellectual capacities of hominids allowed them to generate many situations for which cooperation paid off. The most important of these was the capacity to make and use tools, and the capacity to generate novel behavioral procedures to achieve a goal. The expanded exploitation of the savannah and woodland niche—made possible by tool use—allowed individuals to acquire food items too large to be consumed by a single individual (Isaac 1978; Tooby and DeVore 1987). This created effective opportunities

[5] For example, TIT FOR TAT is an ESS if, and only if, the probability that two individuals will meet again is greater than the larger of these two numbers: $(T\text{-}R)/(T\text{-}P)$ and $(T\text{-}R)/(R\text{-}S)$ (Axelrod 1984).

to provide large benefits to other individuals at a low cost to oneself. There is virtually no cost to sharing food that you cannot consume anyway, and tomorrow you may be the one who has found no food. Fossil evidence suggests that Pleistocene hunter-gatherers, like their modern counterparts, engaged in extensive food-sharing (e.g., Isaac 1978). Similarly, the cost of sharing tools is low compared to the benefits one can garner through using them—and the cost of sharing *information* about tool making may be even lower.

When combined with their capacity to opportunistically manipulate the environment through tool use, our ancestors' ability to generate novel behavioral procedures[6] created situations in which coordinated, cooperative behavior could produce vast payoffs. Perhaps one of the best examples is the "profits" to be made through cooperative hunting. Acting together, several armed men can bring down a large game animal; acting alone, a single armed man is more likely to fail.

These conditions set the stage for the coevolution of a tightly interwoven complex of adaptations whose mutual reinforcement made cooperation more and more advantageous (Tooby and DeVore 1987, and references therein). Cooperative hunting provided a compact and nutritious food source that provided an efficient means for males to invest in offspring; the increased payoff of this pattern of male parental investment favored the restructuring of mating relationships associated with paternity certainty, potentially moderating intra-band male–male competition; such increased parental investment allows larger brains and longer periods for maturation and learning; these, in turn, allow more efficient cooperation and tool use, which leads to even more nutritious food sources from both hunting and gathering; these developments allow the allocation of an increasing proportion of metabolic resources to brain tissues. Such developments may be mutually reinforcing because the evolution of sophisticated cooperation in a large array of activities depends upon a complex cognitive base.

Reconstruction of the causal sequences that fed the evolution of cooperation (and associated developments) is still a matter of debate (cf. Kinzey 1987). The most important point is that the Pleistocene hunter-gatherer environment in which we evolved provided many opportunities for individuals to benefit from mutual cooperation, and cooperation for mutual benefit is a pervasive and inextricable aspect of all past and modern human cultures.

The peculiarities of hominid behavioral ecology place some species-specific constraints on a computational theory of social exchange in humans. Exchange in most primates is restricted to relatively few "items": food, sexual access, defense, grooming. The fewer the items for exchange, the

[6] An ability that some other primates also possess, to a lesser extent. For example, de Waal (1982) shows pictures of chimpanzees who have discovered that they can get past an electrified fence surrounding a tree with edible leaves. One chimpansee holds a large branch against the tree as a ladder, while another climbs it into the tree. The chimpanzee in the tree then throws juicy leaves down to his compatriots on the ground.

more "item-specific" the algorithms regulating exchange can (and should) be: What counts as "error"—cheating or underreciprocating—can be more closely defined, increasing the accuracy of one's mental accounting system and the accuracy of reference (see Part 3). In contrast, because of the human penetration of the "cognitive niche" (Tooby and DeVore 1987), human algorithms for regulating social exchange should be able to handle a wide and ever-changing array of "items" for exchange: tools, participation in coalitional aggression, information about tool-making, participation in opportunistically created, coordinated behavioral routines. This suggests that our algorithms for regulating social exchange, and the associated cognitive capacities that they require to function properly, will have some human-specific properties. These will be discussed in Part 3.

## PART 3. A COMPUTATIONAL THEORY OF SOCIAL EXCHANGE

David Marr has argued that the first and most important step in understanding an information-processing problem is developing a "theory of the computation" (Marr 1982; Marr and Nishihara 1978). This theory defines the nature of the problem to be solved; in so doing, it allows one to predict properties that any algorithm capable of solving the problem must have. Computational theories incorporate "valid constraints on the way the world is structured—constraints that provide sufficient information to allow the processing to succeed" (Marr and Nishihara 1978, p. 41).

For humans, an evolved species, natural selection in a particular ecological situation defines and constitutes "valid constraints on the way the world is structured" for a particular adaptive information processing problem (Cosmides and Tooby 1987). In the case of social exchange, the ecological and game-theoretic aspects of hominid social exchange discussed above provide the ingredients for the construction of just such a computational theory. The ability to engage in a possible strategy of social exchange presupposes the ability to solve a number of information processing problems that are *highly specialized.* The elucidation of these information processing problems constitutes a computational theory of social exchange. Any psychological theory purporting to account for the fact that humans are able to engage in social exchange must be powerful enough to realize this computational theory—that is, its information processing mechanisms must produce behavior that respects the constraints imposed by the selective process. Thus, it must be powerful enough to 1) permit the realization of a "possible" social exchange strategy, i.e., a strategy that can be selected for, and 2) exclude "impossible" strategies, i.e., strategies would be selected against.

The problems most specific to social exchange will be incorporated into a "grammar of social contracts" in the second half of Part 3. A grammar of social contracts is the set of assumptions about the rules governing a par-

ticular social exchange that must somehow be incarnated in the psychological mechanisms of both participants. However, the grammar of social contracts does not exhaust the set of information processing problems posed by social exchange. The ability to successfully participate in social exchange also requires a number of other, associated cognitive capacities, some of which are necessary in a wide range of other evolutionarily crucial social interactions, such as mating, pair-bonding, parenting, and aggression. Before we progress to the grammar of social contracts, five associated cognitive capacities entailed by social exchange will be examined:

1. The ability to recognize many different individuals.
2. The ability to remember aspects of one's history of interaction with different individuals.
3. The ability to communicate one's values to others.
4. The ability to model the values of other individuals.
5. The ability to view as costs and benefits items that one perceives as causally connected to biologically significant variables; human algorithms regulating social exchange should not be too closely tied to particular items of exchange.

Undoubtedly, a clever programmer could design many different algorithms capable of solving these problems. It is even possible that one or two of them could be solved, albeit less efficiently, by domain general mechanisms such as associative nets. But to demonstrate that such mechanisms could, in theory, solve these problems would be to miss the point. The point of using natural selection theory in creating computational theories is that it allows you to specify a set of problems that humans ought to be able to solve quickly, reliably, efficiently, and without explicit instruction. These are problems for which natural selection should have produced specialized, domain specific Darwinian algorithms: *modules* in Marr's (1982) or Fodor's (1983) terminology, *mental organs* or *cognitive competences* in Chomsky's (1975) terminology, *adaptations* in the terminology of evolutionary biology. It is the presumption that natural selection has designed psychological mechanisms that are particularly good at solving these problems that carries implications for the study of attention, communication, the organization of memory, implicit inference, and learning. We shall briefly sketch a few of these implications, together with some of the relevant data.

## HUMAN SOCIAL EXCHANGE REQUIRES SOME FUNDAMENTAL COGNITIVE CAPACITIES

### Proposition 1. One Must be Able to Recognize Many Different Individual Humans

The basic idea is that an individual must not be able to get away with defecting without the other individuals being able to retaliate effectively.

> The response requires that the defecting individual not be lost in a sea of anonymous others. (Axelrod and Hamilton 1981)

Individual recognition is important even if one has an exchange relationship with only one individual. It is that much more important if one has such relationships with a number of individuals; the ability to cooperate with more than one individual is particularly useful to a hunter-gatherer. In order to cooperate only with individuals who are likely to reciprocate, and avoid (or cheat on) individuals who are likely to cheat, one must be able to discriminate different individuals.[7] One need not rely on "preliminary hunches" (Carey and Diamond 1980, p. 60) in singling out individual recognition as a domain for which humans ought to have specialized mechanisms; it is a direct prediction of an evolutionary perspective.

Indeed, humans do seem to have a highly developed ability to recognize large numbers of different individuals. Recognition rates are over 90% for familiar faces that have not been seen for up to 34 years (Bahrick et al. 1975). Patients with a lesion in a specific part of the right hemisphere develop a selective deficit in their ability to recognize faces, called prosopagnosia (Gardner 1974). Carey and Diamond (1980) present and review an impressive array of evidence from a wide variety of sources suggesting that humans have innately specified face-encoding schemas. We are also good at identifying individuals solely through recognizing their idiosyncratic gaits (Cutting et al. 1978; Kozlowski and Cutting 1977).

## Proposition 2. One Must be Able to Remember Some Aspects of the Histories of One's Interactions with Different Individuals

First, one must be able to recognize that a previous interactant in a social exchange is, in fact, a particular previous interactant, and not a stranger or a misidentified individual. Second, once an individual has been identified as a previous interactant, information regarding that individual's particular history of previous interactions, coded in terms of whether that individual has been a cooperator or a cheater, must become accessible to the decision procedures. Third, one needs an "accounting system" for keeping track of who owes who what. As discussed in Part 1, many Pleistocene social exchanges involved "reciprocal altruism"—exchanges in which reciprocation was delayed, not simultaneous. In a simultaneous, face-to-face exchange, if you see that the other person has come prepared to defect, you simply withhold what that person wants.[8] However, the capacity for en-

[7] Organisms that lack the ability to recognize different individuals can also evolve a limited ability to cooperate, but only because of ecological restrictions on their interactions to a very few partners with whom they are in constant and/or exclusive physical proximity (Axelrod and Hamilton 1981).

[8] One would expect people to assume, in the absence of information to the contrary, that such intercontingent behavior occurs in face-to-face interactions. They should be more likely to suspect someone of intending to cheat in delayed benefit transactions.

gaging in transactions in which reciprocation is delayed requires a mental accounting system for keeping track of who owes who what (note: Proposition 5, about item-independent representations, also applies to this accounting system).

The extent of the history of interaction that must become available to the decision procedures regulating participation in social exchange (and whether any of these facts need be consciously recalled) will depend on the details of the particular decision procedures humans have evolved. For example, TIT FOR TAT (Axelrod and Hamilton 1981) requires only that last transaction with each interactant be recalled. But TIT FOR TAT is a successful strategy in a highly constrained and uniform universe where all transactions are simultaneous, the same payoff matrix applies to each transaction, and the size of the payoffs for both players is equal within each transaction (Axelrod 1984). In contrast, payoff matrices in the real world are always in flux, and part of that flux is caused by the negotiative skills of the individuals involved. Moreover, violence is possible in the real world: Exchange situations with individuals who can reliably use violence to get their way do not necessarily fit the constraints of a Prisoner's Dilemma (Tooby and Cosmides, in preparation, a). Thus, an algorithm better adapted to conditions in the real world might assess many more factors regarding one's past history with an individual, such as 1) the number of transactions one has had with that individual in the past, 2) how he or she behaved those transactions (reputation), 3) the size of payoffs to both parties in previous transactions, 4) whether his or her tendency to cheat varied with the size of the payoff involved, 5) whether the conditions governing his or her tendency to cheat have been shifting over time, 6) his or her aggressive formidability, 7) how likely one is to meet that individual in the future (e.g., one party is moving away or likely to die soon), and 8) whether one has accepted a past benefit but has not reciprocated yet. Information regarding others' histories of reciprocation, including circumstances and dispositions that might explain their occasional (or systematic) failures to reciprocate, should be intensely interesting, as should information about how others regard oneself. The power of single defections to be socially communicated through reputation may provide an especially intense disincentive to cheat and hence may have allowed far more reliable systems of cooperation to develop in humans than among species whose ability to communicate such information is limited.

A decision procedure that used such data, current behavioral cues,[9] and the payoff matrix for the current interaction to compute the conditional

[9] For example, a person's facial expression might telegraph his or her intention to cheat. All else equal, a person's "likeability" should be a function of his or her tendency to reciprocate, and cues that suggest "good intentions" ought to be judged more likeable (e.g., sneers and aggressive scowls do not suggest good intent). Although other explanations are possible, it is interesting that people remember unfamiliar faces better when, during initial inspection, they are asked to judge the person's likeability than when they are asked to assign sex (Carey and Diamond 1980).

probability that one's partner will cooperate, might be better adapted to the complexities of exchange in nature.[10] If so, then the need to take such factors into account has implications regarding the organization of human memory. Information about one's history of interaction with a particular person ought to be "filed" with that person's "identity" and activated quickly and effortlessly when an exchange-relevant situation with that person arises. When the payoff matrix of the current interaction is such that "you" will lose a great deal if "I" cheat you, then more of our past exchange history should become accessible than for trivial exchanges. When you believe that I have cheated you in a major way, there should be a flood of memories about your past history with me: You must decide whether it is worth your while to continue our relationship. In addition, this information will help you negotiate with me if you choose to continue our relationship: You can communicate how large a cost I have inflicted on you now and in the past (so I can make amends if I want to continue the relationship), tell me how close you came to ending our relationship (i.e., categorizing me as a permanent defector), convince me that I have become increasingly untrustworthy, threaten to injure my reputation by telling others about my past transgressions, and so on.

The activation of past situations in which I have cheated you may, in turn, activate other[11] affective mechanisms that communicate cost/benefit information: They may cause you to cry, turn your back on me, scream at me, or hit me. The extent and nature of the overt aspects of your affective reaction communicates to me your view of the extent of my injury of you: whether you view it as serious enough to require restitution, how much is required and how soon, whether you intend to cut me off if I defect again, etc. Emotion communication can be viewed as one way individuals communicate costs, benefits, and behavioral intentions to others in negotiative situations (see Cosmides 1983).

## Proposition 3. One Must be Able to Communicate One's Values to Others

To engage in an exchange with you, I must know what you want. Although language is certainly a useful means for communicating what one values, nonlinguistic organisms can also engage in social exchange—however, the

[10] An algorithm was submitted to Axelrod's computer tournament that computed the conditional probability that an interactant would cooperate based on whether that individual had cooperated or defected in past interactions (REVISED DOWNING). It cooperated only when this conditional probability was greater than 50% (random). Its downfall was that it did not discount past behavior relative to present behavior. Therefore, it was exploited by certain programs that became more likely to cheat in later interactions. In a sense, it failed because it assumed that competitor programs had static "personalities."

[11] We say "other" because we see no principled way of drawing a dividing line between emotion and cognition. The flood of memories commonly experienced when a person is betrayed is as much a part of one's "emotional reaction" as turning red and attacking (see Tooby 1985; Tooby and Cosmides, in preparation, b).

range of items they can exchange is necessarily more limited. For example, chimpanzees recruit support from others in aggressive encounters and frequently form long-term coalitional relationships (e.g., de Waal 1982). These coalitions are social exchanges in which the exchanged "item" is mutual aid in fights. A chimpanzee under attack bares its teeth, emits a fear scream, looks at the individual from whom it wants support, and holds out its hand, palm up, toward that individual. If the attacked chimpanzee receives the requested support, its demeanor changes radically: Its hair stands on end, it emits aggressive barks, and it charges its opponent—looking over its shoulder frequently to see if its supporter is still with it. If the chimpanzee does not receive support, it continues cowering with hair flat and teeth bared, screaming and holding out its hand to solicit support.

One also must be able to communicate dissatisfaction with a defector. This also can be done without language, as is vividly illustrated by an interaction between Puist and Luit, two chimpanzees in the Arnhem chimpanzee colony in the Netherlands. Puist and Luit had a long-standing coalitional relationship: Puist had a long history of aiding Luit whenever he attacked or was under attack, and Luit had a long history of extending similar aid to Puist.

> This happened once after Puist had supported Luit in chasing Nikkie [another chimpanzee]. When Nikkie later displayed [aggressively] at Puist she turned to Luit and held out her hand to him in search of support. Luit, however, did nothing to protect her against Nikkie's attack. Immediately Puist turned on Luit, barking furiously, chased him across the enclosure and even hit him. (de Waal 1982, p. 207)

The communication of desires, entitlements, and unfulfilled obligations is possible without language, given that the communicators are both programmed to understand the signals. It requires that a gestural/referential system be shared by the potential cooperators.

A cognitive system that can enable the communication of desires requires more than the development of a few signs. The signs must be coupled with a *referential* system. If I want to exchange an axe for something, how do I indicate what I want? Let's say I point to the pear you are holding in your hand. What am I referring to by pointing to the pear? Do I want that particular pear? Any pear at all? Five bushels of pears? A fruit of some kind, not necessarily a pear? To be led to the site where you found such good pears? Do I want you to hold a branch-ladder so I can climb into a tree that has pears? Or a tree with some other kind of fruit? Do I want to use my axe to core the pear, in exchange for half the pear? And so on.

The ambiguity of reference in the absence of a shared referential system is no mere philosophical puzzle (e.g., Quine 1969; Gleitman and Wanner 1982). For example, it is not clear that the infliction of pain, in the absence of a shared referential framework, could communicate what it is that the individual inflicting the pain wants the other individual to stop doing. The difficulty of communicating desires in the absence of a shared system of

reference is illustrated by certain "communication gaps" that occur between two different, but closely related, species of baboons: hamadryas baboons and savannah baboons.

A male hamadryas baboon acquires a "harem" of females by kidnaping juvenile females from other troops. He leads them to water holes and feeding grounds that are widely scattered in the inhospitable Ethiopian badlands. To keep a kidnaped female from straying, the male bites her whenever she wanders even a few feet from where he wants her. But how does the female know what this bite refers to, what it is that the male does not want her to do? This may seem like a straightforward case of "narrowing hypotheses" through conditioning. However, the same herding technique does not work on a female savannah baboon. When she is abducted, the hamadryas male tries to keep her in line by biting her, to no avail. The savannah female never "gets" what it is he wants, and simply runs off. For males, knowing that one can condition hamadryas females by biting them appears to be no more "implicit in the situation" than knowing what a bite means. Savannah-hamadryas hybrid males who live among hamadryas baboons cannot keep a harem—the hybrid male never "figures out" that he can herd females through biting (Hrdy 1981).

Apparently, the learning mechanisms of hamadryas and savannah baboons include different referential systems. Hamadryas males and females both "know" that a bite means "stay with the herd"; savannah baboons do not. The ability to smile, hug, or inflict pain is not enough. A gestural system for indicating preference that is not cognitively coupled to a referential system would be inaccurate at best, and impossible at worst.

The gestural/referential system that allows members of nonlinguistic species to signal costs, benefits, and behavioral intentions to conspecifics can be thought of as an emotion communication system. Indeed, ethologists have traditionally considered such signaling the primary function of emotional expression, studying intention movements, courtship dances, agonistic displays, and aggressive interactions in mammals, birds, reptiles, fish, and insects. Like modern nonhuman primates, our prelinguistic hominid ancestors undoubtedly had such a system and used it to communicate about social exchange. For example, to this day, humans all over the globe share the same facial expressions of emotion (Eibl-Eibesfeldt 1975; Ekman 1982); we even share many of these facial expressions with nonhuman primates (Jolly 1972, pp. 158–159). The same is true for certain auditory signals, like screaming and crying (Eibl-Eibesfeldt 1975). We can think of no reason why the appearance of language would cause this more ancient system to be selected out. Moreover, to the extent that such signals are universally shared, they have some interesting properties that spoken language lacks:

1. Because they are universally shared, emotion signals can be recognized by anyone. By aiding "translation," such signals expand the range of possible interactants to individuals who speak a different language, in-

dividuals who cannot yet speak a language (small children), and distant individuals beyond the reach of speech but not of sight.

2. Emotion signals can function like intersubjective metrics, permitting an observer to scale the values of the person emitting the signal: A very loud scream indicates a greater cost to the screamer than a moderately loud scream. Signals like screams, smiles, and trembling are "analog": The louder the scream, the wider the smile, the more noticeable the tremble—the more strongly the person can be presumed to feel about the situation causing her to scream, smile, or tremble. Words do not provide such convenient indicators of magnitude, precisely because they are arbitrary and discrete symbols. Verbal expressions indicating size of cost or benefit are more "digital": One might reasonably use "very much" to describe the degree of one's desire in both these sentences: "I want very much for my child's cancer to go into remission" and "I want that apple very much"—yet in these two cases the degree of desire is, presumably, vastly different.
3. Emotion signals allow the incidental communication of values to potential interactants. By observing "your" emotional reactions to various situations, even though they are not directed at me, "I" can learn what you value, and hence what sort of exchange you are likely to agree to (see Proposition 4). The verbal alternative is a process akin to writing to Santa Claus: Reciting a long list stating one's preference hierarchy, with periodic updates.[12]

However, the very properties that make a natural language a poor medium for communicating intensity of affect make it an excellent system for indicating "items" of exchange. The variety of "items" that can be exchanged is severely limited in a species that uses only emotion signals. Primates appear to exchange primarily acts of aggression, protection, food, sex, alarm, and grooming. The use of language does not, of course, eliminate the problem of ambiguous reference. In the absence of a shared referential semantics, knowing what a word refers to is no less problematic than knowing what a gesture refers to.[13] But a natural language permits a potentially infinite number of arbitrary, discriminable symbols to be attached to a potentially infinite number of discriminable classes or entities. As new situations arise, new words can be opportunistically created to refer to them. Consequently, language permits a range and specificity of reference impossible in the purely gestural systems of most primates.

[12] Actually, a list stating that you want *X*, *Y*, and *Z* is not sufficient. Your preferences, including items you already have, would have to be hierarchically ordered using some sort of interval scale or indifference curves, because the salient issue is: What would you be willing to *give up* in order to get *X*, *Y*, and *Z*?

[13] This problem has prompted developmental psycholinguists to posit that children have innately specified "hypotheses" about what sorts of entities are likely to have words attached to them. When coupled with articulated models of the world, this hypothesis and model system amounts to a referential semantics (Gleitman and Wanner 1982).

This property of language opens the vast realm of human adaptations associated with planning and tool-use to social exchange. Tool technology continually changes,[14] with new tools being invented constantly. New technologies enable new and constantly changing opportunities for coordinated, cooperative behaviors that can themselves become "items" of exchange. Great benefits can be had by exchanging tools and by participating in the complex and opportunistically shifting cooperative enterprises these allow—*but only if the tools and behavioral routines can be named.* The expanded power of reference that language affords in social exchange may have been one factor selecting for its evolution. It is not clear that any but the simplest tool-using cooperative enterprises could be accomplished with a nonlinguistic gestural system—routines like the chimpanzees' ladder expedition (see footnote 6), which are discovered quite publicly in the context of an emotionally salient event[15] and don't require long periods of planning.

The evolution of language does not obviate the ability to communicate cost/benefit information through emotion signals. In fact, the more items that members of a species can name and exchange, and the more the instrumental value of these items varies between individuals and over time, the more one needs an "item-independent" yet universally understood system for communicating how much one values an item.

Because the variety of items exchanged by nonlinguistic primates is limited, some sets of items whose marginal value did not vary could (hypothetically) have unique cost/benefit weightings associated with them that are shared by most other members of the species (e.g., ten grooms deserves one assist in a fight, a season of protection by a male deserves exclusive sexual access at the height of estrus, etc.). Theoretically, such items could have a preprogrammed, universally recognized "exchange rate."

But there can be no preprogrammed, universally acknowledged "exchange rate" for a constantly changing array of tools and coordinated behavioral routines. Language combined with emotion signaling affords a uniquely powerful communicative system for social exchange in a planning, tool using, and opportunistically cooperative species. A wide variety of items can be precisely specified through language, and their relative value to an individual can be simultaneously communicated, either incidentally[16] or in-

[14] At least for *Homo sapiens sapiens*. The *Homo erectus* tool kit appears surprisingly constant over a wide range of different environments, from Asia to Africa, for over 1 million years (Pilbeam, pers. comm.). Of course, this observation applies only to tools that are recognizable as such in the fossil record. For example, a branch used as a ladder would not show up in the fossil record.

[15] The Arnhem chimpanzees discovered the ladder trick when one screaming chimpanzee, fleeing from a very public attack, bounded up a broken branch that happened to be resting against a tree.

[16] Because the incidental communication of cost/benefit information is important (see Proposition 4), one might predict that, all else equal, individuals are more likely to emit emotion signals in the presence (or suspected presence) of potential reciprocators than when they are alone. Similarly, they should be more likely to suppress emotion signals in the presence of potential aggressors—value information helps aggressors; it tells them what they should threaten to kill, destroy, or prevent.

tentionally, via emotion signals. Indeed, there is preliminary evidence suggesting that some aspects of the acoustic expression of emotion in humans have been integrated into our species-specific language capacity in ways that facilitate the communication of values and intentions (Cosmides 1983).

## Proposition 4. One Must be Able to Model the Values of Other Individuals

In some ways, Proposition 4 is just the other side of Proposition 3: One must have a cognitive system capable of decoding communications of the sort described in Proposition 3. In addition to this, however, one ought to have learning mechanisms that are specialized for picking up incidental information about the values of potential interactants—for doing "marketing research." In order to propose an exchange for mutual benefit, one must have some notion of what kind of "item" the other individual is likely to value. The individual who is well-equipped to do "marketing research" on potential interactants will be able to initiate far more exchanges than the individual who waits for potential interactants to intentionally announce their preference hierarchies.

Because emotion signals flag cost/benefit information, they should automatically recruit attention and be difficult to ignore. An ear-splitting scream should be more difficult to ignore than an equally loud train whistle; soft sobbing from the next room should be harder to ignore than the loud honk of a car horn outside. A broad smile should recruit more attention than the sound of a motor starting up or than waves in tall grass as it is blown by the wind.[17] Attention should be more sustained for emotion signals emitted by a potential interactant: The cry of a friend should recruit more sustained attention than the cry of a stranger.

Not only should attention be drawn to emotion signals, but one's learning mechanisms should be quick to pick up what the signal refers to—what, exactly, the person emitting the signal is reacting to. This implies that our referential semantics (see footnote 13) includes "hypotheses" about what kinds of events emotion signals are likely to refer to—hypotheses about what other individuals are likely to value. Having such hypotheses is all the more important because many negative emotion signals refer to valued items that are not present or have not happened, vastly complicating the task of assigning a referent. When a person is hungry, he or she may moan because the thing valued—food—is *not* present. Others must infer the desire for food from the moan, even though there is no spatio-temporally contiguous event in which the signal (moan) and the referent (food) are both present.

Evolutionary theory provides a rich heuristic base for developing theories about what kinds of preference information are included in our ref-

[17] Conditioned stimuli linked to events producing large costs or benefits should also recruit attention, e.g., a fire engine siren on your street.

erential semantics. Because humans are tool users, planners, and cooperators who can invent many alternative means for realizing a particular goal, many specific items of human preference will differ from culture to culture in ways that depend on that culture's technology, ecology, social system, and history. This does not mean, however, that desires are random. Evolutionary theory is rife with hypotheses regarding what states of affairs the typical human is likely to prefer (see Cosmides 1985, pp. 165–167). A cognitive "list" of typical human preferences would still be inadequate, however, because there are complex interactions between competing preferences that evolutionary theory speaks to (e.g., what do you do if your spouse beats you, but he is your only source of income?). Therefore, the algorithms that guide our "marketing research" must include cost/benefit analysis procedures that allow one to take such complexities into account in modeling other people's values.

Although researchers from Bartlett (1932) to Schank and Abelson (1977) have posited that pragmatic inference is guided by "schemas," "frames," or "scripts"—domain specific inference procedures—they have provided little insight into their specific content. Using evolutionary biology as a guide, the system so far proposed (default hypotheses about typical human preference hierarchies plus procedures for combining factors) provides a starting place for elucidating the content of "motivation scripts"—algorithms that guide pragmatic inference about human preference and motivation.

Motivation scripts should be powerful and sophisticated, for the ability to model other people's values is useful in a wide variety of evolutionarily important social contexts, from social exchange to aggressive threat to mate choice to parenting. They should prove to be strong organizational factors in the construction and reconstruction of memories. Details that are normally considered insignificant should be more easily recalled when activated motivation scripts allow them to be perceived as causally linked to biologically significant variables.[18] Veridical recall of stories that violate the assumptions about human preference instantiated in our motivation scripts should be difficult (as, indeed it is: e.g., Bartlett 1932). Motivation scripts should guide the reconstruction of such stories during recall, distorting the original story in ways that make motivational sense. Implicit motivational assumptions are so pervasive in human communication that motivation scripts will probably be an essential component of any artificial intelligence program that can usefully converse in a natural language.

An emotion signal should not only recruit attention and activate one's own motivation scripts, it should arouse one's curiosity. One would expect increased tendencies to observe the emotion-arousing event and ask ques-

[18] Owens, Bower, and Black (1979) present evidence of this kind. Interestingly, the most biologically significant motivational theme (an unwanted pregnancy) elicited the highest recall of mundane details about a character's day.

tions about it. Crowds gather around fights, children follow fire trucks to the scene of a fire, onlookers bombard police with questions at the scene of a crime. Journalists make a profession of gathering information about the values and behavior of people who have a large (real or perceived) impact on our lives. Motivation scripts may guide inferences about what exactly a given emotion signal refers to, but it can do this only if it is fed concrete information. The concrete information one acquires by witnessing an emotion-arousing event fills in parameter values in motivation scripts, determining which data structures and inference procedures are appropriate in decoding the reacting person's values.[19]

Acquiring information about the values of potential interactants is, in itself, valuable. Decoding the value systems of potential interactants is therefore likely to become a cooperative enterprise in itself. We even have a name for such exchanges of information and "analysis"—gossip. Gossip is usually about situations that cause emotional reactions in potential interactants—exactly the kind of situations that provide a window into someone's values. The more biologically significant the information, the "hotter" the gossip: Events involving sex, pregnancy, fights, windfalls, and death should be particularly "hot" topics, especially when they signal a change in someone's needs, values, or capacity to confer benefits. Hot gossip should be particularly interesting and easily remembered. Gossip about people who can have a large impact on one's well-being should be especially interesting; gossip about people one does not know should be comparatively boring. Similarly, cues or indicators of the character of potential interactants, including their disposition to cheat or defect, are themselves extremely valuable information, differentially attended to and exchanged. Reputation is an important kind of social information, and it plays a significant role in the social life of any stable group of humans.

The learning mechanisms that guide such "marketing research" should produce *person-specific models* of the preferences and motivations of potential and actual interactants. General motivation scripts help build person-specific preference models; these become more elaborated the more contact one has with that particular person. As this happens, inferences drawn from a person-specific model will generate more accurate interpretations of that person's behavior and emotion signals than inferences drawn from the general motivation scripts.

It would be useless for information about the preferences of different individuals to be stored together in a semantic network, filed under "preferences" or "values." Like information about an individual's history of reciprocation, a model of an individual's preferences and motivations should

[19] There are, of course, other good reasons for being curious about biologically significant events, e.g., you yourself might be confronted with the same situation at some point. However, when such events impact potential interactants, they should be *especially* interesting: A fist fight in your academic department provokes more interest than one among strangers in another city.

be "filed" under his or her identity. When the opportunity to acquire more preference information about an individual arises, the model appropriate to that individual must be easily retrieved, not just a model of average preferences. "Averaging" the fact that one person prefers *Z* to *W* but another person prefers *W* to *Z* into one model of "average" preference does not enhance one's ability to engage in social exchange.[20] In contrast, learning that "Smith values W more than *X* more than *Y* more than *Z*" and that "Jones values *Z* more than *X* more than *Y* more than *W*" increases your ability to make offers that benefit you *given the limits imposed by what Smith or Jones are willing to accept*. Offering *W* to Smith is more likely to induce him to give you *Y* than offering him *Z*; exactly the reverse is true of Jones. If you value *Z* more than *W*, you are better off making Smith an offer; if you value *W* more than *Z*, then strike a deal with Jones. The proper decision can be made only if person-specific preference information can be conveniently retrieved.

## Proposition 5. Human Algorithms Regulating Social Exchange Should Not be Closely Tied to Particular Items of Exchange

That tools, information about tool making, and participation in opportunistically created, coordinated behavioral routines were important items for exchange has implications for the structure of human cognitive algorithms regulating social exchange. The more limited the range of items exchanged, the more specific the algorithms regulating exchange can be. For example, the items exchanged in a cleaning fish symbiosis can be directly specified in the algorithms regulating the exchange. The host fish is specifically programmed to discriminate cleaner fish from similar-looking prey items, and, upon recognizing one, to refrain from eating it. The cleaner fish is specifically programmed to discriminate a host fish from other large, predatory fish, and, upon recognizing one, to approach and eat its ectoparasites (Trivers 1971). Whereas the exchange algorithms of other organisms can be specific to the relatively few items they exchange, human algorithms regulating social exchange should be able to take a wide variety of input items, as long as these items are *perceived* as costs and benefits to the individuals involved in the exchange.

However, despite our remoteness from the Pleistocene and the range of items commonly traded then, some items should be more readily perceived as costs and benefits: those for which the perceiver can ascertain a clear causal link to adaptively significant variables like offspring, kin, sex, food, safety, shelter, protection, aggressive formidability, and dominance. Evidence for this is so universal that there seems little point in belaboring it.

[20] Although noting that most people in your culture prefer *W* to *Z* might enhance your ability to recognize and participate in social exchanges with new interactants. One might expect such culture-specific information to be incorporated into the "typical human" motivation scripts.

For example, a Mr. Michael Pastore of Dallas recently made the following comment in an interview in the *Wall Street Journal*:

> "I never pay for dinner with anything other than my [American Express] Platinum Card when I'm on a first date," says the 30-year-old seafood importer, flashing his plastic sliver inside the glitzy Acapulco Bar. "Women are really attracted to the success that my card represents." (The *Wall Street Journal*, April 17, 1985, p. 35)

Mr. Pastore perceives a clear causal link between his "plastic sliver" and a biologically significant variable: the ability to attract sexual partners. His perception that a Platinum Card can attract sexual partners is based in turn on the perception that owning one is causally linked to a variable that is biologically significant to females in choosing male sexual partners—the ability to accrue resources.[21] Knowing this, one can infer that Mr. Pastore perceives owning an American Express Platinum Card as a *benefit*, and that if he did not own one, he might well be willing to give up other items in order to acquire one. It is a suitable item for social exchange.

The prediction, then, is that the algorithms regulating social exchange in humans will be *item-independent*. Furthermore, they will operate on *cost-benefit representations* of the interaction. As we will argue in the next section, any interaction that is interpreted as having a particular, characteristic, cost-benefit structure will be categorized as an instance of social exchange and will call up procedural knowledge specialized for reasoning about this domain.

## THE GRAMMAR OF SOCIAL CONTRACTS

A grammar of social contracts specifies the properties that must be embodied by Darwinian algorithms for reasoning about social exchange. It incorporates the strategic constraints outlined in Part 1 and the ecological constraints outlined in Part 2.

Just as a grammar of the English language is a set of rules for distinguishing well-formed sentences from ill-formed sentences, a grammar of social contracts is a set of rules for distinguishing well-formed social contracts from ill-formed social contracts. It includes the set of assumptions about the rules governing social exchange that must somehow be incarnated in the psychological mechanisms of both participants. Without these assumptions, much of what people say, mean, and intend to do in exchange situations could not be understood or anticipated. This grammar creates the "cohesion of discourse" (Wason and Johnson-Laird 1972, p. 92), and the cohesion of behavior, in interactions involving uncoerced exchange. It con-

[21] In fact, cross-cultural evidence is accumulating that indicates that a potential mate's ability to accrue resources is more important to women than to men, just as evolutionary theory predicts (Buss 1987).

stitutes the procedural knowledge that individuals must share in order to communicate their intentions to others in this particular kind of negotiative interaction (see Cosmides 1983).

Moreover, in establishing what conditions must hold if a social contract is to be recognized as well-formed, the grammar also provides a framework for understanding what deviations from these conditions mean. For example, certain types of deviations will indicate "dishonorable" intentions (see discussion of "baseline fraud" below), others will indicate an insult (e.g., "I'd pay a whole nickel to sleep with your mother"), and still others will indicate that the speaker is either joking (e.g., "I'd sell my firstborn for a cigarette") or eccentric ("I'll give you my paycheck for your gumwrapper"). The grammar of social contracts thus provides a framework for understanding the many shades and colors of meaning in situations involving exchange. The ability to formally specify such "grammars," and hence analyze conformity and systematic deviations from them, provides a basis for assimilating the "interpretation of cultures" (Geertz 1973) into a cogent evolutionary framework. Cultural systems of meaning have been widely regarded among social anthropologists as being entirely beyond the capability of an evolutionary perspective to address or explain (Sahlins 1976). However, because such "symbolic productions" are both produced and interpreted by adaptively structured cognitive mechanisms, "interpretation" can be converted from a literary exercise to an exercise in evolutionary cognitive psychology. As an increasing number of adaptive "grammars" of cognitive mechanisms are mapped, the nature and systematic properties of cultural meaning systems may become increasingly apparent, and tractable to an evolutionary and empirical treatment.

We shall describe the grammar of social contracts in pedantic detail, because it serves several important empirical and theoretical functions. First, it is a productive source of hypotheses for uncovering the cognitive mechanisms involved. Second, such explicitness is required to rigorously construct a cognitive theory, e.g., one specified enough to implement in an artificial intelligence program. The creation of such an AI program is the test of whether one really has a complete cognitive theory, and at a minimum, any AI program capable of understanding and reasoning about social exchange would have to embody this grammar of social contracts. This exercise is useful because it exposes the hidden assumptions and complexities that theoretical handwaving misses. Because we have such innate algorithms, these tasks seem transparently simple, and this illusion of simplicity constantly leads researchers to underestimate the intricacy of the cognitive machinery necessary to perform these tasks. And third, no psychological theory claiming to account for how people learn to engage in social exchange can be considered adequate unless it can be shown that the learning mechanism proposed is powerful enough to permit the acquisition of the grammar of social contracts. Thus, a well-specified computational theory of social ex-

change provides a crucial test of adequacy that any proposed psychological theory must be able to pass (Cosmides and Tooby 1987).

The items valued by our hominid ancestors were correlated with costs and benefits in their inclusive fitness; otherwise social exchange could not have evolved. The strategic exigencies of exchanging items that had real effects on the inclusive fitness of the exchangers selected for algorithms programmed with a particular set of cost/benefit relations (see Fig. 1, Part 1). These relations can be expected to regulate how we think about social exchange, even if the items we value today are no longer correlated with our inclusive fitness. The grammar of social contracts specifies these cost/benefit relations.

Unlike the exchange algorithms of cleaner fish or even baboons, human algorithms for regulating social exchange should be item-independent: They should represent items of exchange as costs and benefits to the participants, and operate on those representations (see Proposition 5). The proposed grammar of social contracts is therefore expressed in cost/benefit terminology.

What must *P* and *Q* stand for if the sentence "If *P* then *Q*" is to instantiate a well-formed social contract?

To make the discussion concrete, let's fill in some values for *P* and *Q* in the offer "If *P* then *Q*." Let's say "I" offer "you" the following contract: "If you walk my dog, then I'll give you a million dollars." *P* stands for "you walk my dog" and *Q* stands for "I'll give you a million dollars." Likewise, not-*P* stands for "you do not walk my dog" and not-*Q* stands for "I do not give you a million dollars."

At the time of this offer, but independent of it, you have a certain level of "well-being" and certain expectations about the future, all of which play some part in determining what you would, at this point, consider to be of value. This baseline will be termed your *zero level utility*. For simplicity's sake, let us assume that 1) value is subjective, and 2) the individual is the final arbiter of what he or she finds valuable. Natural selection theory has much to say about what kinds of items and states most humans will consider valuable (i.e., about preferences and motivations; see Propositions 4 and 5), but such considerations are unnecessary for this analysis.

## What Conditions Must Hold for You to Accept This Offer?

Let us consider what conditions must hold for you to accept this offer. Your zero level utility baseline is derived from a vast number of conditions and expectations about the state of the world. In the absence of this offer, one of those expectations about the future must be not-*Q*—you do not expect to be receiving $1,000,000 from me. If not-*Q* comes to pass, your utility level will not have moved from your zero level baseline, 0(you).

*Q*—receiving $1,000,000 from me—must be something that you consider to be a benefit. An "item"—an act, entity, or state of affairs—is a

*benefit to you* (*B*(you)) if, and only if, it increases your utility above your zero level baseline.[22] Assuming you value having a million dollars (Q) more than you value not having a million dollars (not-*Q*), then *Q*—having a million dollars—constitutes a benefit to you. You will not accept this offer unless, at the time of acceptance, you believe that *Q* constitutes a benefit to you. Using terms defined with respect to your values (rather than mine), we can rephrase my offer as: "If *P* then *B*(you)."

An item is a *cost to you* (*C*(you)) if, and only if, it decreases your utility below your zero level baseline.[23] In this offer, *P*—walking my dog—is the item that I have made my offer of *B*(you) contingent upon. *Usually, P* will be something that you would not do in the absence of an inducement; otherwise, I would be incurring a cost (giving up *Q*, the million dollars) by making the offer (if you were going to walk my dog anyway it would be foolish of me to offer you the million dollars).[24] If *P* is *not* something you expected to do in the absence of my offer, then, in your value system, not-*P* (not walking my dog) is part of your zero level baseline, 0(you). This means that if not-*P* comes to pass, you will not have moved from your zero utility baseline, and you will be no worse off than if my offer had never been made. If we posit that my dog is ugly and vicious, and that walking him would embarrass you, endanger your health, and assault your aesthetic sensibilities, then *P* (walking my dog) decreases your utility and is therefore a cost to you, *C*(you).

Stated in terms of *your* value system, my offer can now be rephrased as "If *C*(you) then *B*(you)." But other conditions must hold before you will accept my offer. There is a constraint on the magnitudes (absolute values) of *B* and *C*, namely, *B*(you) > *C*(you), or, equivalently, *B*(you) minus *C*(you) > 0. For you to accept my offer, a million dollars must be more of a benefit to you than walking my ugly dog is a cost. If this is not the case there would be no point in your entering into the contract; it would not increase your utility. The greater the magnitude of *B* minus *C*, the more attractive the contract will appear. A contract that reversed this constraint (such that $C >> B$) sounds perverse. For example, the following offer strikes people as foolish: "If you break your arm then I'll give you a penny." Unsurprisingly, Fillenbaum (1976) found that subjects consider such offers "extraordinary" 75% of the time, compared to a 13% rate for offers that fit the constraints described above.

[22] Presumably there are costs and benefits associated with any action. More precisely, *B*(you) is a net benefit—the benefit to you of receiving $1 million is greater than the cost to you of receiving $1 million.

[23] Again, this is a net cost—the cost to you of walking my dog is greater than the benefit to you of walking my dog.

[24] *P* does not have to be a *C*(you) for you to accept my contract, although I must *believe* that it is a *C*(you) in order to offer the contract in the first place. You could be trying to defraud me into offering this contract by dissembling about your real intentions. Perhaps you have been planning all along to walk my dog, but led me to believe that you are not planning to walk it so I would make you an offer. See below, on baseline fraud.

## What Conditions Must Hold For Me to be Willing To Make an Offer?

We can also consider the contract from the point of view of the person offering it, for simplicity's sake identified as "me." What conditions must hold for me to be willing to offer a contract? First, I must believe that not-*P* (your not walking my dog) will come to pass if I do not make the offer. This means that not-*P* is a component of my zero level baseline: If not-*P* comes to pass, my utility level will not have changed. Second, I must want *P*—in my value system, having my dog walked must increase my utility, it must be a *benefit to me* (*B*(me)). Third, not-*Q*—not giving you $1,000,000—*usually* will be part of my zero level baseline, 0(me); if you do not accept my offer, I do not plan on giving you $1,000,000, and if not-*Q* comes to pass, I will not have moved from my zero utility baseline.[25] Fourth, if not-*Q* is part of my zero baseline, then *Q*—giving you $1,000,000—represents a decrease in my utility and is therefore a *cost to me* (*C*(me)). Fifth, like you, I will not enter into the contract (offer it in the first place) unless *B*(me) > *C*(me) (unless having my dog walked is worth *more* to me than relinquishing the million dollars; this example sounds eccentric precisely because most readers would assume that it violates this constraint).

In other words, I want *P*, and I am willing to give up *Q* to get you to do *P*; but I am not willing to give up *Q* without getting *P*. (I want you to walk my dog and I am willing to give up $1,000,000 to get you to walk him; but I am not willing to give up my $1,000,000 without your walking him.)

In your value system, "If *P* then *Q*" translates to: "If *C*(you) then *B*(you)." (i.e., "If I incur the cost of walking her dog, then I will get the benefit of receiving $1,000,000 from her.") However, in my value system the same offer translates to "If *B*(me) then *C*(me)" (i.e., "If I get the benefit of your walking my dog, then I will incur the cost of relinquishing my $1,000,000 to you"). As you can see, *P* represents a different utility level to me (*B*(me)) than it does to you (*C*(you)). The same holds for *Q*. In a well-formed social contract, that is, a contract that I am willing to offer and you are willing to accept, the utility levels associated with *P* and *Q* are those shown in Table 1.

An offer is not entirely symmetrical, however. Suppose there were some way of equating value systems. Although $B > C$ for both of us (or else we would not both agree to the contract), *P* (walking my dog) might be a smaller cost to you than *Q* (giving up $1,000,000 to you) is to me (or vice versa). Likewise, *Q* might be a larger benefit to you than *P* is to me. These asymmetries may lead to a difference in the magnitude of our "profit margins" (*B* minus *C*). Where it is possible to vary the effective magnitude of the acts

[25] *Not-Q* being part of my zero level baseline is not a *necessary* condition for my making an offer, but it is necessary that you *believe* it is part of my zero baseline if you are to accept my offer. Unknown to you, I might intend to give you $1,000,000 regardless, but want to get as much as I can in return.

Table 1. Cost/Benefit Translation of My Offer into Your Value System and Mine

My offer: "If *P* then *Q*"
("If you walk my dog then I'll give you $1m")

| | | Your Point of View | My Point of View |
|---|---|---|---|
| *P* | (you walk my dog) | *C*(you) | *B*(me) |
| *not-P* | (you do not walk my dog) | 0(you) | 0(me) |
| *Q* | (I give you my $1m) | *B*(you) | *C*(me) |
| *not-Q* | (I do not give you my $1m) | 0(you) | 0(me) |

or items of exchange (together with their associated costs and benefits), unequal profit margins invite bargaining: Each may attempt to increase his or her "profit margin" at the expense of the other. Bargaining under such constraints constitutes an antagonistic game, because more *B*(you) per unit *C*(you) corresponds to more *C*(me) per unit *B*(me). (See Fig. 2; for a fuller account of these intercontingent relations and their psychological sequelae, see Tooby 1975). However, as long as your profit margin is greater than zero, it is in your interest to accept my offer, *regardless of how large my profit margin is* (and vice versa). If $B > C$ for both parties, then both parties have benefited from the exchange. For this reason, we consider the term "subtle cheating," which Trivers (1971) uses to describe an interaction in which profit margins are unequal, to be a misnomer. "Under-reciprocating" is a more appropriate term; "cheating" is more usefully reserved for the violation of a contract.

## Baseline Fraud

There is a joke that runs like this:

> A man from out of town walks up to a woman and says "If you sleep with me three times I'll give you $15,000." She is hard up for cash, so she agrees. After each session he pays her the money he promised. The woman decides this is an easy way to make money, so after she has been paid the full $15,000 she asks him if he would like to continue the arrangement. He says he can't because he must return home the next day. She asks "Where's home?" "Oshkosh," he replies. "Oh!" she says, "That's where my mother lives!" He answers, "Yes, I know. She gave me $15,000 to deliver to you."

The man in the joke has defrauded the woman by concealing information about their zero level baselines.

A contract has been *sincerely* offered and sincerely accepted when each party believes that the $B > C$ constraint holds for the other, in this case, when the contract has the following cost/benefit structure:

Man's offer: "If you sleep with me three times then I'll give you $15,000"
"If *P* then *Q*"
Woman's point of view: "If *C*(woman) then *B*(woman)"
Man's point of view: "If *B*(man) then *C*(man)"

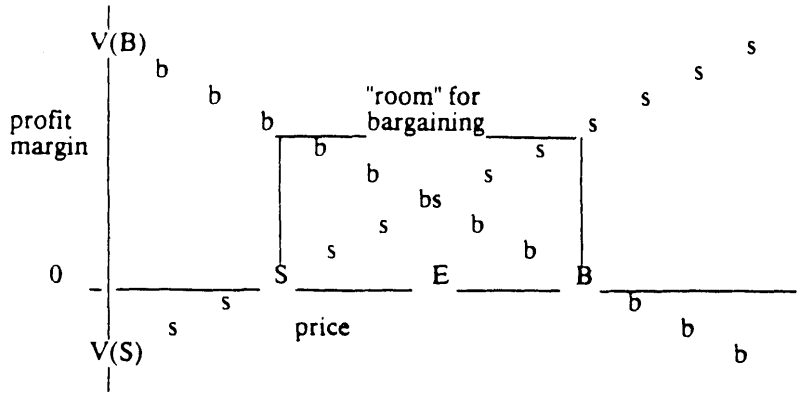


**FIGURE 2.** Negotiation in social exchange (Adapted from Tooby 1975).

Imagine a buyer and a seller haggling over the price of a car. $V(B)$ represents the value of the car to the buyer; if the buyer could get the car for free ($C$(buyer) = price = 0) then $B$(buyer) = $V(B)$, the car's intrinsic value to the buyer. The b-b-b line shows how the potential buyer's profit margin changes as a function of price; the higher the price he pays, the lower his profit margin ($V(B)$ − price). $B$, the point where this line intersects the $x$-axis, is the buyer's breakeven point, the price at which his profit margin is zero. The buyer makes a profit if he pays any price *less* than $B$. $V(S)$ represents the value of the car to the seller; if the seller gives it away ($B$(seller) = price = 0), then $C$(seller) = $V(S)$, the car's intrinsic value to the seller. The s-s-s line shows how the seller's profit margin changes as a function of price; the higher the price he gets, the higher his profit margin (price − $V(S)$). $S$ is the seller's breakeven point, the price at which his profit margin is zero. The seller makes a profit if he sells the car at any price *greater* than $S$. Both buyer and seller profit if the car is sold at any price such that $S <$ price $< B$. They only profit *equally*, however, at price $E$, the point where b-b-b intersects s-s-s. The buyer will try to push the price down the s-s-s curve to S, the seller will try to push the price up the b-b-b-b curve to $B$. The price range between $S$ and $B$ represents "room for bargaining." The buyer might try to convince the seller that the seller's curve is actually steeper and the buyer's shallower, that B is really less than it is (i.e., he threatens to withdraw his offer at a price lower than $B$), that the seller "ought" to give him a break, etc. (and vice versa).

The emotional language of Marxist economics and labor negotiations can be understood with this graph. The worker (person selling his labor) claims he is being "exploited" and that management is earning "excess profits" when the price of an hour of his labor is $S \leq$ price $< E$ (management's "excess profit" is the difference between their profit margin at the price they are currently paying the "exploited" worker and their lower profit margin at $B$, the price the worker prefers). Management (person buying labor) complains that labor unions are strangling the company when workers succeed in pushing the price of labor up such that $E <$ price $\leq B$ ("strangling" implies that price $> B$, a situation that cannot be true if the company is making a profit greater than zero). In truth, both labor and management benefit compared to their other options at any price between $S$ and $B$.

---

The woman in the joke assumed that the man's offer fit these requirements, that he offered a sincere contract. However, the man knew that if the woman knew what *he* knew about her baseline and his, they would both see the structure of the contract as:

| | "If $P$ then $Q$" |
|---|---|
| Woman's point of view: | "If $C$(woman) then 0(woman)" |

Man's point of view: "If *B*(man) then 0(man)"

In actuality, the man gave up nothing in exchange for *B*(man).

Humor is frequently based on the violation of implicit assumptions. The punch line of this joke violates the woman's (and the listener's) implicit assumption that the man had offered a "sincere" contract. Although not-*Q* is *usually* part of the offerer's zero level utility baseline, this is not a *necessary* condition for his making an offer (see above). However, it is a necessary condition of the woman's acceptance that she *believe* that, in the absence of the offer, not-*Q* would come to pass; otherwise, accepting the offer would decrease her utility by *C*(woman).

For any proffered contract of the form: "If you do *P* then I'll do *Q*," the acceptor has been the victim of baseline fraud when:

1. The acceptor believes that not-*Q* will come to pass if he or she turns down the contract, and
2. This belief is false, and
3. The offerer knows the acceptor holds this false belief, and
4. The offerer either fosters the acceptor's false belief, or does nothing to disabuse the acceptor of this belief.

Likewise, the offerer has been the victim of baseline fraud when:

1. The offerer believes that not-*P* will come to pass if he or she does not offer the contract (or if it is turned down), and
2. This belief is false, and
3. The acceptor knows the offerer holds this false belief, and
4. The acceptor either fosters the offerer's false belief, or does nothing to disabuse the offerer of this belief.

Had the woman wanted to sleep with the man all along, regardless of payment, she would have thought she was tricking him by getting the added benefit of $15,000 (until she heard about her mother's gift). This is because the offerer's belief that the potential acceptor's zero level baseline includes not-*P* (not sleeping with him) is a necessary condition for the offerer to make the offer, but it is not a necessary condition for the acceptor to accept the offer. Baseline fraud is different from cheating: In baseline fraud both parties have, technically, honored their contractual obligations. As will be seen, this is not the case with cheating.

## Summary of the Structure of Sincere Social Contracts

The conditions that hold when an individual sincerely offers or sincerely accepts a social contract are shown in Table 2. For the sake of simplicity, *P* and *Q* stand for the actual items exchanged (these can be actions as well as entities). The first column shows the contract's cost/benefit structure in terms of the sincere offerer's value system; the second column shows what the sincere offerer believes the contract's structure is in terms of the ac-

Table 2. Sincere Social Contracts: Cost/Benefit Relations When One Party is Sincere and That Party Believes the Other Party is Also Sincere

| | My offer: "If you give me *P* then I'll give you *Q*" | | | |
|---|---|---|---|---|
| | Sincere Offer | | Sincere Acceptance | |
| | I Believe: | | You believe: | |
| *P* | *B*(me) | *C*(you) | *B*(me) | *C*(you) |
| *not-P* | 0(me) | 0(you) | 0(me) | 0(you) |
| *Q* | *C*(me) | *B*(you) | *C*(me) | *B*(you) |
| *not-Q* | 0(me) | 0(you) | 0(me) | 0(you) |
| *Profit margin* | Positive: *B*(me) > *C*(me) | Positive: *B*(you) > *C*(you) | Positive: *B*(me) > *C*(me) | Positive: *B*(you) > *C*(you) |
| *Translation* | | | | |
| *My terms* | "If *B*(me) then *C*(me)" | | "If *B*(me) then *C*(me)" | |
| *Your terms* | "If *C*(you) then *B*(you)" | | "If *C*(you) then *B*(you)" | |

ceptor's value system. The third column shows the contract's cost/benefit structure in terms of the sincere acceptor's value system; the fourth column shows what the sincere acceptor believes the contract's structure is in terms of the offerer's value system. The table shows that the sincere offerer and the sincere acceptor view the contract's cost/benefit structure in exactly the same way.

Table 3 shows what conditions hold when one person offers or accepts a contract sincerely, but is the victim of baseline fraud. The sincere person believes the contract fits the conditions specified in Table 2. However, the defrauder believes the contract fits the criteria specified in Table 3. Furthermore, if the sincere person were to find out that he or she had been tricked concerning baseline information, that person would share the defrauder's view of the contract's cost/benefit structure. An analysis of baseline fraud is useful because it serves to distinguish the conditions that *must*

Table 3. Baseline Fraud: Cost/Benefit Relations When a Sincere Party Makes a Social Contract with an Individual Perpetrating a Baseline Fraud

| | My offer: "If you give me *P* then I'll give you *Q*" | | | |
|---|---|---|---|---|
| | I try to defraud you; you accept sincerely If you knew what I knew, we would both believe | | You try to defraud me, I offer sincerely If I knew what you knew, we would both believe | |
| *P* | *B*(me) | *C*(you) | 0(you) | 0(me) |
| *not-P* | 0(me) | 0(you) | ? | *C*(me) |
| *Q* | 0(me) | 0(you) | *B*(you) | *C*(me) |
| *not-Q* | ? | *C*(you) | 0(you) | 0(me) |
| *Profit margin* | Positive: *B*(me) > *C*(me) | Negative: *C*(you) | Positive: *B*(you) > *C*(you) | Negative: *C*(me) |
| *Translation* | | | | |
| *My terms* | "If *B*(me) then 0(me)" | | "If 0(me) then *C*(me)" | |
| *Your terms* | "If *C*(you then 0(you)" | | "If 0(you) then *B*(you)" | |

hold for a contract to be offered or accepted, from those that need not hold, but usually do.

That people represent actions as costs and benefits with reference to a zero point based on their current expectations is a psychological prediction that is not strictly necessitated by natural selection theory in its simplest form. However, reciprocation theory does require that the individual realize a net increase in its fitness from participation; this could, in principle, be computed using an ordinal preference scale without reference to a zero point.

We use this system because we believe it provides a powerful means by which the individual can distinguish exchanges from other kinds of intercontingent behavior. For example, most people would probably recognize the utterance "If you call the police, then I'll shoot you" as a threat. Yet 1) it has the same linguistic form as a contract—"If *P* then *Q*," and 2) like the person who accepts a sincere contract, the person threatened will realize an increase in fitness by obeying the threat instead of defying it.

However, the hypothesis that humans represent events as costs and benefits with respect to a zero utility point based on their current baseline expectations provides a straightforward means of distinguishing threats from contracts. A contract has the form "If *C*(you) then *B*(you)." However a threat has the form "If 0(you) then *C*(you)." In the absence of the threat, the hearer in the example intends to call the police; it is part of his or her zero level baseline. Being shot is not in the hearer's plans; it constitutes a cost. This representational system allows the principled differentiation of the various forms of intercontingent behavior. We hypothesize that recognition of the form of intercontingent behavior at hand (social exchange, threat, etc.) automatically activates the set of rules appropriate for reasoning about it. In the next section, we sketch the rules appropriate to social exchange.

## Social Contracts as "Speech Acts"

The relations specified in the previous sections are implicit in the sincere offer of a contract and its sincere acceptance. But to understand cheating (a violation of the contract), we have to analyze what contractual obligations "you" and "I" incur by entering into a contract. This calls for a brief foray into "speech act" theory (Searle 1971).

Speech act theory is a part of analytic philosophy that grew out of the realization that, in speaking, people frequently do more than simply refer to something in the world. Frequently they *do* something by virtue of saying something. When I say "I promise to *X*", for example, I am not referring to something in the world: I am *making* a promise, and thereby incurring certain obligations—I have committed a "speech act" (e.g., Searle 1971). "Offering a contract" and "accepting a contract" can both be considered speech acts. Thus, we can ask the question: What do I mean when I say "If you give me *P* then I'll give you *Q*" and what do you mean when you say

you "accept" my offer? Grice (1957, 1967) has provided a convenient structure for understanding the meaning of speech acts.

In committing a speech act,

> something [a behavior, intention, or frame of mind] intentionally is produced in another with the intention that he realize why it was produced and that he realize he was intended to realize all this (Nozick 1981, pp. 369–370, on Grice).

Using this structure and the cost/benefit analysis above, when an actor, "I", offers a contract by saying, "If you give me *P* then I'll give you *Q*," the actor means:

1. I want you to give me P,
2. My offer fulfills the cost/benefit requirements of a sincere contract (listed in Table 2).
3. I realize, and I intend that you realize, that 4–9 are entailed if, and only if, you accept my offer:
4. If you give me *P*, then I will give you *Q*,
5. By virtue of my adhering to the conditions of this contract, my belief that you have given (or will give) me *P* will be the cause of my giving you *Q*,
6. If you do not give me *P*, I will not give you *Q*,
7. By virtue of my adhering to the conditions of this contract, my belief that you have not given (or will not give) me *P* will be the cause of my not giving you *Q*,
8. If you accept *Q* from me, then you are obligated to give me *P* (alternatively, If you accept *Q* from me then I am entitled to receive *P* from you),
9. If you give me *P*, then I am obligated to give you *Q* (alternatively, If you give me *P* then you are entitled to receive *Q* from me).

These rules capture the intercontingent nature of social exchange: They specify the ways in which the behavior of one person is contingent upon the behavior of another person. Some philosophical refinements are discussed in Appendix 1. However, these points are not essential to the rest of the article.

Offering a contract is somewhat more complicated than other speech acts (like promises; see Searle 1971) in that none of the conditions apply unless the hearer *accepts* the contract. In contrast, the conditions for a promise (or a threat) hold regardless of whether the hearer consents. Making a promise is a unilateral act; making a contract is not. In saying that one accepts an offer, the acceptor means that he or she understands, and agrees to comply with, the conditions specified in 1–9 (above).

On first inspection it might seem that a contract actually expresses a biconditional: "*Q* if and only if *P*." If this were the case, the terms of the contract would be violated (someone would have cheated) if you are not in possession of *Q* after having done *P* (I cheated you), or if you are in possession of *Q* without having done *P* (you cheated me). But it is not a bi-

conditional because a social contract involves the twin notions of obligation and entitlement.

What does it mean for you to be obligated to do *P*?

1. You have agreed to do P for me under certain contractual conditions (like 1–9), and
2. Those conditions have been met, and
3. By virtue of your not thereupon doing *P*, you agree that if I use some means of getting *P* (or its equivalent) from you that does not involve getting your voluntary consent, then I will suffer no reprisal from you.

Alternatively, 3 can be:

3. By virtue of your not thereupon giving me *P*, you agree that if I lower your utility by some (optimal) amount *X* (where $X > B$(you)—your unearned gains), then I will suffer no reprisal from you.

The first formulation expresses restitution, the second, punishment. One would expect the tendency to punish to be greatest when restitution is not possible. To our knowledge, the conditions determining the optimal size of *X* have not yet been formally analyzed. We suspect the optimal *X* would be large enough to deter future cheating but small enough that it does not discourage future cooperation. However, it is clear that a cheater would not be deterred by an *X less than or equal to* $B$(cheater). With $X = B$(cheater), the potential cheater will be indifferent between cheating and cooperating; with $X < B$(cheater) the potential cheater will realize a net benefit by cheating.

To take reprisal against someone trying to claim "just" restitution or punishment is to indicate that you are no longer interested in continuing a relationship with that person. In the contretemps between Puist and Luit, the two chimpanzees discussed in Proposition 3, Luit *allowed* Puist to punish him for his defection. The judgment that this punishment was "allowed" can be made because Luit is far stronger than Puist, and in a direct test of strength Puist would not have a chance against Luit (de Waal 1982). To do otherwise would have signaled an end to their several year reciprocal relationship.

What does it mean for "you" to be entitled to *Q*?

1. I have agreed to give you *Q* under certain contractual conditions (like 1–9), and
2. Those conditions have been met, and
3. By virtue of my not thereupon giving you *Q*, I agree that if you use some means of getting *Q* (or its equivalent) from me that does not involve getting my voluntary consent, then you will suffer no reprisal from me.

As in obligation, an alternative formulation of 3 is:

3. By virtue of my not thereupon giving you *Q*, I agree that if you lower my utility by some (optimal) amount *X* (where $X > B$(me)—my unearned spoils), then you will suffer no reprisal from me.

Thus, the notions of entitlement and obligation are reciprocally related: My being entitled to receive *P* from you is equivalent to your being obligated to give me *P* and vice versa.

A social contract is not a biconditional because "I" must do that which I am obligated to do, but I am not required to accept that to which I am entitled. If I pay the cost that I am obligated to pay (*C*(me), which corresponds to *B*(you)), I have fulfilled my end of the contract; I do not have to accept the benefit (*B*(me)) that I am entitled to (however, you must offer it). Unless conditions have changed, failure to accept a benefit one is entitled to is foolish (and rare—such behavior would have been strongly selected against), but no one regards it as violating the terms of the contract. Only selectional thinking makes sense of these otherwise arbitrary features of human social cognition.

## Looking for Cheaters

As discussed, under biologically plausible circumstances, indiscriminate cooperation cannot be selected for. The game-theoretic structure of the natural selection process dictates that social exchange can evolve only if it is governed by a strategy that requires reciprocation and excludes or retaliates against cheaters. To implement such a strategy, human social contract algorithms must include procedures that allow us to quickly and effectively infer whether someone has cheated, or intends to cheat, on a social contract. For a mechanism to detect cheating, it must have a specification of what conditions constitute cheating.

Cheating is the violation of the conditions of a social contract. It is the failure to pay a cost to which you have obligated yourself by accepting a benefit. The social contract can be explicit or implicit,[26] a private agreement or a rule of one's social group.

Let's assume "I" offered, and "you" accepted, the following contract: "If you give me *P* then I'll give you *Q*." In your value system this translates to: "If *C*(you) then *B*(you)." You have cheated me when you have accepted the item that corresponds to *B*(you) (item Q) without giving me the item that corresponds to *C*(you) (item *P*). In other words, you have cheated me when you have accepted item *Q* from me, but you have not given me item *P*. This means I have paid *C*(me) (item *Q*), but have not received *B*(me) (item *P*). Your payoff: *B*(you). My payoff: *C*(me).

In my value system, the same contract translates to: "If *B*(me) then *C*(me)." I have cheated you when I have accepted *B*(me) (item *P*) without paying *C*(me) (item *Q*). In other words, I have cheated you when I have

[26] Given that hominids probably participated in social exchange long before they had language, one would expect the act of accepting a benefit to frequently be interpreted as implicit agreement to a social contract—as a signal that the acceptor feels obligated to reciprocate in the future. (Of course, one would expect the donor to jump to this interpretation more readily than the acceptor!) This view is formalized in British and U.S. contract law—a contract is invalid unless some "consideration" has changed hands–even a symbolic $1 will suffice.

Table 4. How Do You and I Make Out When One of Us Cheats the Other?

| | I Cheat You | | You Cheat Me | | Contract Fulfilled | |
|---|---|---|---|---|---|---|
| You give me $P$ | $B$(me) | $C$(you) | — | — | $B$(me) | $C$(you) |
| You do not give me $P$ | — | — | 0(me) | 0(you) | — | — |
| I give you $Q$ | — | — | $C$(me) | $B$(you) | $C$(me) | $B$(you) |
| I do not give you $Q$ | 0(me) | 0(you) | — | — | — | — |
| My payoff | $B$(me) | | $C$(me) | | $B$(me)–$C$(me) | |
| Your payoff | $C$(you) | | $B$(you) | | $B$(you)–$C$(you) | |

accepted item $P$ from you, but have not given you item $Q$. This means you have paid $C$(you) (item $P$), but have not received $B$(you) (item $Q$). Your payoff: $C$(you). My payoff: $B$(me). These relations are summarized in Table 4.

As mentioned in Proposition 5, social contract algorithms in humans must represent items of exchange as costs and benefits to the participants, and operate on those representations. The detection of cheating depends on modeling the exchange's cost/benefit structure from the point of view of one's partner, as well as from one's own point of view. Thus, for any given exchange, two descriptions of each item must be computed by the social contract algorithms. For a sincere contract, "If you give me $P$, then I'll give you $Q$," item $P$ should be described as both $B$(me) and $C$(you), and item $Q$ should be described as both $C$(me) and $B$(you) (see Table 4). The cost/benefit structure to oneself should be easily recoverable, even if the contract is phrased in terms of the value system of one's exchange partner.[27] There is a structural parallel to transformational grammars as they were initially conceptualized: The "surface structure" is the way the offer is actually phrased; the deep structure is a cost/benefit description of the surface structure from the point of view of each participant. The deep structure of the offer incorporates the information shown in Table 2 (or 3, if one party is "baseline defrauding"). A prediction of this computational analysis is that these cost/benefit structures are the descriptions from which participants construct paraphrases and reconstruct the course of the interaction from memory.

Inference procedures for detecting cheaters must operate on a cost/benefit description of the contract from the potential cheater's point of view. These procedures must allow one to quickly and effectively infer that individual $X$ has cheated when one sees that $X$ has accepted $B(X)$ but not paid $C(X)$. When a transaction has not yet been completed, or when one's information about a transaction is incomplete, "look for cheaters" procedures should lead one to:

1. Ignore individual $X$ if $X$ has *not* accepted $B(X)$.
2. Ignore individual $X$ if $X$ has paid $C(X)$.

[27] However, one might predict that an offer phrased in terms of the potential acceptor's value system might sound more attractive, indicating that the offerer *really* understands (has a good model of) what the potential acceptor wants.

3. Watch out for individual $X$ if $X$ has accepted $B(X)$.
4. Watch out for individual $X$ if $X$ has *not* paid $C(X)$.

In situations 1 and 2, individual $X$ cannot possibly have cheated; in situations 3 and 4, individual $X$ can cheat. One keeps an eye on $X$ in situation 3 to make sure $X$ fulfills his or her obligation by paying $C(X)$. One keeps an eye on $X$ in situation 4 to make sure $X$ does not abscond with $B(X)$, to which he or she is not entitled.

Most people would not guess that the structure of a simple, straightforward social exchange is as complex as described here. But then, this illusion is a prediction of the theory. People usually do not realize how complex the grammar of their language is, yet they produce grammatical sentences with ease. Similarly, people do not realize how complex engaging in social exchange is, yet they do it with ease. Both parties implicitly understand and act on all the relations involved because both possess the same Darwinian algorithms for reasoning about social exchange. These algorithms perform these computations reliably and automatically, thereby shielding us from awareness of the underlying complexity, and leaving us with the impression that what it takes to perform a social exchange is simple and self-evident—an activity that does not require any extraordinary computations.

## PART 4. EMPIRICAL EVIDENCE SUGGESTS THAT HUMANS DO, IN FACT, HAVE A "LOOK FOR CHEATERS" PROCEDURE

Whether the human cognitive architecture contains an array of special purpose, domain specific, procedure-rich modules, or consists entirely of a few, major, domain general information-processing mechanisms, is still at issue in modern cognitive science. To date, this debate (involving such issues as learnability, innateness, and so on) has been conducted primarily within the fields of psycholinguistics and perception, and has left most other subfields of cognition largely untouched. Human reasoning, especially, has traditionally been considered domain general: The innate processes hypothesized—whether "logical," "inductive," or associationistic—have been thought of as operating uniformly, regardless of content, with content-dependent performance attributed to the vagaries of differential experience.

However, the evolutionary approach to cognition makes very different predictions. Such an approach predicts that when a category of content is thematically linked to important recurrent adaptive problems, specialized procedures for dealing with that category of content will be invoked, and they will operate on that content differently than on other categories of content. Social exchange is a domain for which the evolutionarily-predicted computational theory is complex, and the fitness costs associated with "errors" are large. Successfully conducted social exchange was such an important and recurrent feature of hominid evolution that selection would have

favored a reliable, efficient cognitive capacity specialized for reasoning about social exchange. A general-purpose learning mechanism that operated on all kinds of content indiscriminately is necessarily more inefficient at specialized problems: Among other things, being initially "ignorant," it will make costly errors and will continue to do so throughout its learning phase. (More tellingly, no one has even been able to propose a general "design" that could "learn" how to conduct social exchange, and it remains an open question whether such a system is possible.) Presumably, the costly trade-offs necessary to make a learning system general would put it at a selective disadvantage, and accordingly a phylogenetically antecedent general learning system presumably would either be supplanted by a domain-specific system when social exchange became significant in the hominid lineage or would be used primarily for learning in other less structured domains.

The computational theory of social exchange discussed above, derived from the evolutionary theory of reciprocation (Trivers 1971; Axelrod and Hamilton 1981), provides specific predictions about the human psyche, including:

1. The human psyche includes specialized cognitive algorithms that govern how people reason about social exchange;
2. These algorithms to operate on item-independent, cost/benefit representations of exchange interactions; and
3. These algorithms include specialized procedures that are efficient at detecting potential cheaters, cheating being defined with respect to the grammar of social contracts (Part 3).

Cosmides (1985) and Cosmides (in press) conducted a series of experiments designed to test for the existence in the human psyche of these hypothesized algorithms, using a test of logical reasoning known as the Wason selection task. These experiments are briefly reviewed below.

In psychology, the study of human reasoning started from the premise that humans reason logically, i.e., in accordance with the rules of inference of the propositional calculus (Wason and Johnson-Laird 1972). These rules of inference are content-independent: they generate only true conclusions from true premises, regardless of what the propositional content of the premises is.

However, more than a decade of research has shown that people rarely reason according to these canons of formal logic. Moreover—and contrary to initial expectations—psychologists found that human reasoning is content-dependent: the subject matter one is asked to reason about seems to regulate how people reason. This can be seen very clearly in experiments using the Wason selection task (Wason 1966), a test of logical reasoning in which one is asked to determine whether a conditional rule has been violated (see Fig. 3). In general, performance on the Wason selection task is very poor, when the standard of "correct" is the propositional calculus. However, the content of a few rules elicits a high percentage of "logical" re-

## *a.* Abstract Problem (AP)

Part of your new clerical job at the local high school is to make sure that student documents have been processed correctly. Your job is to make sure the documents conform to the following alphanumeric rule:

"If a person has a 'D' rating, then his documents must be marked code '3'."
(If P then Q )*

You suspect the secretary you replaced did not categorize the students' documents correctly. The cards below have information about the documents of four people who are enrolled at this high school. Each card represents one person. One side of a card tells a person's letter rating and the other side of the card tells that person's number code.

Indicate only those card(s) you definitely need to turn over to see if the documents of any of these people violate this rule.


## *b.* Drinking Age Problem (DAP; adapted from Griggs & Cox, 1982)

In its crackdown against drunk drivers, Massachusetts law enforcement officials are revoking liquor licenses left and right. You are a bouncer in a Boston bar, and you'll lose your job unless you enforce the following law:

"If a person is drinking beer, then he must be over 20 years old."
(If P then Q )

The cards below have information about four people sitting at a table in your bar. Each card represents one person. One side of a card tells what a person is drinking and the other side of the card tells that person's age.

Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking this law.

**c. Structure of Social Contract (SC) Problems**

It is your job to enforce the following law:

Rule 1 — Standard Social Contract (STD-SC): "If you take the benefit, then you pay the cost."
(If P then Q)

Rule 2 — Switched Social Contract (SWC-SC): "If you pay the cost, 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 paid the cost.

Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking this law.

| | Benefit Accepted | Benefit NOT Accepted | Cost Paid | Cost NOT Paid |
|---|---|---|---|---|
| Rule 1 — STD-SC: | (P) | (not-P) | (Q) | (not-Q) |
| Rule 2 — SWC-SC: | (Q) | (not-Q) | (P) | (not-P) |

**FIGURE 3.** Content effects on the Wason selection task. The logical structures of these three Wason selection tasks are identical; they differ only in propositional content. Regardless of content, the logical solution to all three problems is the same: to see if the rule has been violated, choose the *P* card (to see if it has a *not-Q* on the back), and choose the *not-Q* card (to see if it has a *P* on the back). This is because only the combination on the same card of a true antecedent (*P*) with a false consequent (*not-Q*) can violate—and thereby falsify—a conditional rule. Therefore, the logically correct response is *P and not-Q*.

Only 4–25% of college students choose both these cards for abstract problems (see panel *a*). The most common responses are *P and Q*, or *P*: subjects rarely see the relevance of the *not-Q* card. Yet about 75% see its relevance for the Drinking Age Problem (panel *b*)—a familiar "standard" social contract (see panel *c*, Rule 1)—and choose both *P* and *not-Q*. Poor performance on the abstract problem is not due solely to the use of "abstract" symbols; similar rates of responding are usually found for a number of more familiar, "thematic" conditionals: relations between food and drink, cities and means of transportation, schools and fields of study.

Panel *c* shows the abstract structure of a social contract problem. A "look for cheaters" procedure would lead a subject to choose the "benefit accepted" card and the "cost not paid" card, *regardless of which logical categories they represent*. For a *standard* social contract, like Rule 1 or the Drinking Age Problem, the correct social contract response—*P and not-Q*—converges with formal logic. However, for a *switched* social contract, like Rule 2, the correct social contract response—*not-P and Q*—diverges from formal logic. According to social contract theory, the Drinking Age Problem reliably elicits logically correct responses because it is a standard social contract, and not because it is familiar. Note: The logical categories (*P*s and *Q*s) marked on the rules and cards(*) are here only for the reader's benefit; they never appear on problems given to subjects.

sponses (for review, see Cosmides 1985). This effect is known as the "content effect" on the Wason selection task.

When the content effect on the Wason selection task was first observed, a number of researchers tried to account for it in terms of differential experience. Although it is difficult to do justice to the richness of these hypotheses briefly, fundamentally they proposed that familiarity (differential experience) with the rule being tested increases the probability that a subject will produce the logically correct response (however, different theorists proposed different mechanisms to account for this phenomenon; see, e.g., Manktelow and Evans 1979; Griggs and Cox 1982; Johnson-Laird 1982; Pollard 1982; Wason 1983). The problem with these hypotheses was that some familiar content seemed to produce the content effect, whereas other familiar content did not.

Cosmides (1985) reinterpreted the existing literature by pointing out that virtually all of those few familiar rules that did produce a robust and replicable content effect happened to have the cost/benefit structure of a social contract, as described in Part 3: They could be translated as "If you take the benefit, then you pay the cost." Moreover, in reasoning about these rules, subjects behaved as if they were "looking for cheaters": They investigated persons who had accepted benefits (to see if they had failed to pay the required cost) and persons who had failed to pay the required cost (to see if they had illicitly absconded with the benefit). For "standard" social contract rules, such as the ones that were tested, the "benefit accepted" card and the "cost not paid" card happen to correspond to the logical categories *P* and *not-Q*, respectively (see Figure 3, panel C, Rule 1). This means that a subject who is looking for cheaters will choose the two cards that correspond to the logically correct response, *P and not-Q, by coincidence*, because of the accidental correspondence between the logical and social contract categories. This accounts for why subjects appeared to reason logically about standard social contract rules, but not about familiar rules that were not social contracts.

However, all the standard social contract rules tested were familiar. To experimentally determine whether the content effect is due to the hypothesized "look for cheaters" procedure, rather than to familiarity, Cosmides tested *unfamiliar* rules that either did or did not correspond to social contracts.

Cosmides took highly unfamiliar rules, such as "If a man eats cassava root, then he has a tattoo on his face," and embedded them in two different contexts. One context transformed the rule into a standard social contract, by telling the subject that cassava root was a rationed benefit and that having a facial tattoo was a cost to be paid. The other context made the rule describe some (non-social contract) aspect of the world (e.g., men with tattoos live in a different place than men without tattoos; cassava root grows only where the men with tattoos live; so maybe men are simply eating foods that are most available to them). If the content effect is due to familiarity, then both

problems should yield low levels of logically correct responses, because both test unfamiliar rules. However, if the content effect is due to the presence of a "look for cheaters" procedure, then the unfamiliar social contract problem should yield high levels of logically correct responses, because for standard social contracts, the "benefit accepted" card and the "cost not paid" card happen to correspond to *P and not-Q*, the logically correct response. This is, in fact, what happened: while only 23% of subjects chose *P and not-Q* for the unfamiliar descriptive problems, 73% made this response to the unfamiliar standard social contracts. Moreover, the *unfamiliar* social contract problems elicited more logically correct responses than *familiar* descriptive problems did (the social contract effect was about 50% larger than the effect that familiarity had on descriptive problems).

These experiments eliminated the hypothesis that familiarity alone can account for the content effect on the Wason selection task. Furthermore, they showed that when a rule has the cost/benefit structure of a social contract, subjects are very good at "looking for cheaters," even when the situation they are reasoning about is unfamiliar and culturally alien.

To eliminate the hypothesis that social contract content somehow enhances logical reasoning, Cosmides next tested unfamiliar social contracts that were *switched*: These are rules that translate to, "If you pay the cost, then you take the benefit." If social contract content causes subjects to reason logically, then they would choose the logically correct response, *P and not-Q*, for switched social contracts, just as they did for standard ones, even though these cards correspond to individuals who could not possibly have cheated (see Fig. 3, panel C, Rule 2). However, if reasoning on social contract rules is guided by a "look for cheaters" procedure, then subjects would choose *not-P and Q*, a response completely at variance with formal logic. This is because for a switched social contract, the "cost not paid" card corresponds to the logical category *not-P*, and the "benefit accepted" card corresponds to the logical category *Q* (see Fig. 3, panel C, Rule 2). A "look for cheaters" procedure should be blind to logical category: It should cause the subject to choose the "benefit accepted" card and the "cost not paid" card, regardless of their logical category, because these are the cards that represent potential cheaters.

This prediction was also confirmed. The switched social contracts elicited the "look for cheaters" response, *not-P and Q*, from 71% of subjects, even though this response is illogical according to the propositional calculus. In comparison, the unfamiliar descriptive problems (i.e., those not depicting a social contract) elicited this illogical response from only 2% of subjects, and elicited the logically falsifying response from 14.5%.

In the above experiments, social contract rules were pitted against descriptive rules. However, in a further set of experiments, social contract rules were pitted against "permission" rules that lacked the cost-benefit structure of a social contract. Cheng and Holyoak (1985) had proposed that the modern individual social experience of permissions causes people

(through an unspecified mechanism employing some kind of induction) to build a cognitive schema that would produce the same pattern of responses we term "the social contract effect." In their theory, permission rules have the format, "If an action is to be taken, then the precondition must be satisfied." To test between the two theories, the rules tested were all given the permission format, but, as before, they were embedded in different contexts. One kind of context portrayed taking the action as a benefit and satisfying the precondition as a cost, thereby transforming the rule into a social contract, whereas the kind of other context portrayed a permission situation, but one that was not a social contract (see Cosmides, in press).

Seventy-five percent of subjects gave the correct social contract answer (standard: *P and not-Q*; switched: *not-P and Q*) in response to the social contract problems, compared to only 21% of subjects for the non-social contract permission problems. This result supports the hypothesis that humans represent social exchange in item-independent, cost/benefit terms. The "look for cheaters" procedure is not activated by prescriptive or "permission" rules that lack an easy or obvious cost/benefit interpretation. In contrast, even a highly unfamiliar rule can activate the "look for cheaters" procedure, as long as the context in which it occurs allows the subject to map the cost/benefit valuations that the participants to the exchange assign to the items mentioned in the rule.

The results of these experiments show that when subjects reason about situations involving social exchange, their responses follow the distinctive "adaptive logic" predicted for the "look for cheaters" procedure (described in Part 3), and that this procedure is activated by item-independent, cost/benefit representations of exchange interactions. The results are inconsistent with alternative theories of human performance on logical reasoning tasks, including theories proposing that humans reason in accordance with formal logic, experience-based associational theories, and induction-based theories.

The finding that adult subjects are very adept at detecting potential "cheaters" on a social contract, even when it is unfamiliar and culturally alien, stands in marked contrast to the repeated finding that they are not skilled at detecting potential violations of descriptive and permission rules, even though such rules are commonly encountered in everyday life. The ontogeny of the algorithms that produce these results remains an open question. It is possible that they are, in some carefully delimited sense, learned. However, the mental processes involved appear to be powerfully structured for social contracts, yet weakly structured for other elements and relations drawn from common experience. This implies that the "look for cheaters" procedure is either itself innate, or else the product of a learning process that is guided and structured by innate algorithms that are specialized for reasoning about social exchange.

## CONCLUSION

Although adaptive participation in social exchange depends upon the correct solution of complex, highly structured, and highly specialized information

processing problems, humans in all cultures and of virtually all ages both understand and successfully participate in such social exchanges with ease. The costs and benefits of participation in social exchange have also constituted an intense selection pressure over a significant fraction of hominid evolutionary history. It is implausible to expect that natural selection would leave learning in such a domain to the vagaries of personal experience processed through some kind of general-purpose learning mechanism. The evolutionary expectation that humans do indeed have adaptively structured social exchange algorithms receives substantial empirical support from experimental investigations of human reasoning that 1) have falsified the competing domain general theories of reasoning performance on the Wason selection task and 2) have confirmed the presence of the evolutionarily predicted complex of design features that are diagnostic of adaptation in this domain. We argue that this study of social exchange provides an example of how evolutionary and cognitive techniques can be combined to elucidate aspects of human culture and the psychological mechanisms that underlie them.

It is worth noting that the kinds of "domain general" theories that were falsified as explanations for performance on the Wason selection task are the same kinds of theories that are more generally advanced to account for the human "capacity" for culture (Sahlins 1976; Geertz 1973). Because no imaginable state of human affairs is forbidden by such domain general views of culture, there is little in the way of cultural phenomena that is inconsistent with such models, and consequently very little that is predicted or illuminated by them. The view that cultures are arbitrary symbolic productions is widely and justly criticized by advocates of evolutionary approaches. But by using evolutionary psychology, it is possible to go further and meet traditional anthropological theories of culture on their own ground.

As the process of identifying and mapping these innate mechanisms proceeds, mechanism by mechanism, the differing domains of culture each can be analyzed through reference to the highly structured information processing algorithms that govern its expression. The "interpretation of cultures" can be changed from a post hoc literary exercise about arbitrary symbolic productions into a principled investigation grounded in the evolved psychology of humans and its systematic impact on the cultures it produces.

We are very indebted to Jerome Barkow, Martin Daly, Irven DeVore, Roger Shepard, Donald Symons, and Margo Wilson for many stimulating discussions of the issues explored in this paper. We are especially grateful to Jerome Barkow, Nicholas Blurton Jones, Michael McGuire, and an anonymous reviewer for their comments on the various drafts, and to Jason Banfield and Lisa Bork for their help with the manuscript. We would also like to thank Jeff Wine and Roger Shepard (and NSF grant no. BNS 85-11685 to Roger Shepard) for their support.

## APPENDIX 1. PHILOSOPHICAL REFINEMENTS (CATEGORIZED BY CLAUSE)

2. In other words, the cost/benefit requirements *do* hold for me and I *believe* that they hold for you. (Note: sincere cost/benefit requirements entail "I

value getting $P$ from you *more* than I value keeping $Q$." so this need not be added as a separate statement.) Clause 2 is an implication of my offer even if the sincere cost/benefit requirements do not hold. After all, baseline defrauders mean their offers to be thought sincere.

3. ". . . and *I intend* that you realize . . ." In other words, I did not make the offer accidentally. My having made the offer is a consequence of the activation of my social contract algorithms (my belief that the contract would result in a net benefit to me is a necessary condition for my making the offer; see discussion of the meaning of "cause" in clause 5). If my social contract algorithms had not been activated, I would not have made the utterance. This is presumed for a contract that is offered verbally—there are virtually no circumstances under which one can *accidentally* utter a sentence. However, for nonlinguistic primate species, one can imagine scenarios in which "gestures" are accidentally produced. For example, in the course of a fight, a chimp is chased up a tree. The tree limb supporting him breaks, causing him to fall with his arm stretched out. An outstretched arm in the context of a fight is usually a request for support. However, this gesture was made accidentally rather than intentionally; it was not made as a consequence of the chimp's social contract algorithms having been activated. Therefore, ". . . I intend that you realize . . ." is not part of the gesture's meaning. The fact that it was "accidentally" produced robs the "gesture" of its meaning as a request for support.

5. My belief that you have given me $P$ cannot cause me to give you $Q$ in just any old way. For example, the following is *not* the sense of causation meant:

> Let's say you own a priceless statue, and I have some very compromising pictures of you that you want destroyed. I keep these pictures in my car. I make the offer "If you give me the statue ($P$), then I'll destroy the pictures ($Q$)." You agree, unaware that I have no intention of destroying the pictures because I want to continue to enrich myself by blackmailing you. We arrange for you to leave the statue at a drop point. I retrieve it, and my *belief* that you have given me this priceless statue makes me so agitated and nervous that I have an accident, and the car blows up, destroying the pictures. I have, in fact, done $Q$, and my belief that you gave me $P$ caused me to give you what you wanted—$Q$—but not in the right sense of "cause." (e.g., Nozick 1981, p. 369)

The correct notion of "cause" refers to the psychological realization of (the algorithm instantiating) this computational theory and the fact that it is guiding my behavior. My belief that you have given me $P$ fills in the parameter value in the algorithm; this triggers the set of procedures *within* the algorithm corresponding to the contract's conditions of satisfaction. Triggering these procedures results in my giving you $Q$. This is the same sense of "cause" as in a computer program: the information that $P$ can cause a computer to do something by virtue of that information's functional relation to various of its procedures. Let's say I have written a program in Basic instantiating all the conditions for making a social contract. The program

then offers—"If you type '*P*' into me then I'll print '*Q*' for you"—and I accept. Part of the program would involve the computer waiting for me to fulfill my obligation, and this part may be written thus:

10 Input "Now give me P"; A$

20 If A$ = P then go to 40

30 go to 10

40 Print "Q"

My typing *P* gives the variable A$ the parameter value *P* (analogous (?) to the computer believing that I have typed '*P*' into it), and this causes the computer to print '*Q*'. The same sense of cause is meant in clause 7.

## REFERENCES


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

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

Bahrick, H.P., Bahrick, P.O., and Wittlinger, R.P. Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology* 104: 54–75, 1975.

Bartlett, F.C. *Remembering: A Study in Experimental and Social Psychology*. Cambridge, U.K.: Cambridge University Press, 1932.

Buss, D. Sex differences in human mate selection criteria: An evolutionary perspective. In *Sociobiology and Psychology*, C. Crawford, D. Krebs, and M. Smith (Eds.). Hillsdale, N.J.: Erlbaum, 1987.

Carey, S., and Diamond, R. Maturational determination of the developmental course of face encoding. In *Biological Studies of Mental Processes*, D. Caplan (Ed.). Cambridge, Mass., The MIT Press, 1980.

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

Chomsky, N. *Reflections on Language*. New York: Random House, 1975.

Cosmides, L. Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology*. 9: 864–881, 1983.

—— Deduction or Darwinian algorithms?: An explanation of the "elusive" content effect on the Wason selection task. Ph.D. diss., Harvard University, University Microfilms, 1985.

—— The logic of social exchange: Has natural selection shaped how humans reason? *Cognition*, in press.

——, and Tooby, J. From evolution to behavior: Evolutionary psychology as the missing link. In *The Latest on the Best: Essays on Evolution and Optimality*, J. Dupre (Ed.). Cambridge, Mass.: The MIT Press, 1987.

Cutting, J.E., Profitt, D.R., and Kozlowski, L.T. A biomechanical invariant for gait perception. *Journal of Experimental Psychology* 4: 357–372, 1978.

Dawkins, R. *The Extended Phenotype*. San Francisco: W.H. Freeman, 1982.

de Waal, F. *Chimpanzee Politics: Power and Sex Among Apes*. New York: Harper and Row, 1982.

Eibl-Eibesfeldt, I. *Ethology: The Biology of Behavior*, (2nd ed.). New York: Holt, Rinehart and Winston, Inc., 1975.

Ekman, P. (Ed.) *Emotion in the Human Face*, 2nd ed., Cambridge, U.K.: Cambridge University Press, 1982.

Fillenbaum, S. Inducements: On the phrasing and logic of conditional promises, threats, and warnings. *Psychological Research* 38: 231–250, 1976.

Fodor, J.A. *Modularity of Mind*. Cambridge, Mass.: MIT Press, 1983.
Gardner, H. *The Shattered Mind*. New York: Random House, 1974.
Geertz, C. *The Interpretation of Cultures*. New York: Basic Books, 1973.
Gleitman, L.R., and Wanner, E. Language acquisition: The state of the state of the art. In *Language Acquisition: The State of the Art*, E. Wanner and L.R. Gleitman (Eds.). Cambridge, U.K.: Cambridge University Press, 1982.
Goodall, J. van Lawick. The behaviour of free-living chimpanzees in the Gombe Stream Reserve. *Animal Behaviour Monograph* 3, 1968.
—— *In the Shadow of Man*. Boston: Houghton-Mifflin, 1971.
Grice, H.P. Meaning. *Philosophical Review* 66: 377–388, 1957.
—— Logic and conversation. Unpublished William James Lectures, Harvard University, 1967.
Griggs, R.A., and Cox, J.R. The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology* 73: 407–420, 1982.
Hall, K., and DeVore, I. Baboon social behavior. In *Primate behavior*, I. DeVore (Ed.). New York: Holt, 1965.
Hamilton, W.D. The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7: 1–52, 1964.
Hrdy, S. Blaffer. *The Woman that Never Evolved*, Cambridge, Mass.: Harvard University Press, 1981.
Issac, G.Ll. The food-sharing behavior of protohuman hominids. *Scientific American* 238: 90–108, 1978.
Johnson-Laird, P.N. Thinking as a skill. *Quarterly Journal of Experimental Psychology* 34A: 1–29, 1982.
Jolly, A. *The Evolution of Primate Behavior*, New York: Macmillan, 1972.
Kinzey, W.G. (Ed.) *Primate Models for the Origin of Human Behavior*, New York: SUNY Press, 1987.
Kozlowski, L.T., and Cutting, J.E. Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychodynamics* 21: 575–580, 1977.
Lee, R.B. and DeVore, I. (Eds.) *Man the Hunter*, Chicago: Aldine.
Luce, R.D., and Raiffa, H. *Games and Decisions: Introduction and Critical survey*, New York: Wiley, 1957.
Manktelow, K.I., and Evans, J. St B.T. Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology* 70: 477–488, 1979.
Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman, 1982.
——, and Nishihara, H.K. Visual information processing: Artificial intelligence and the sensorium of sight. *Technology Review*: 28–49, October 1978.
Maynard Smith, J. 1982. *Evolution and the Theory of Games*, Cambridge, U.K.: Cambridge University Press.
McCracken, G.F., and Bradbury, J.W. Social organization and kinship in the polygynous bat *Phyllostomus hastatus*. *Behavioral Ecology and Sociobiology* 8: 11–34, 1981.
Nozick, R. *Philosophical Explanations*, Cambridge, Mass.: Harvard University Press, 1981.
Owens, J., Bower, G.H., and Black, J.B. The "soap opera" effect in story recall. *Memory and Cognition* 7: 185–191, 1979.
Pilbeam, David, Department of Anthropology, Harvard University, personal communication.
Pollard, P. Human reasoning: Some possible effects of availability. *Cognition* 10: 65–96, 1982.
Quine, W.V.O. *Ontological Relativity and Other Essays*, New York: Columbia University Press, 1969.
Sahlins, M. *The Use and Abuse of Biology*, Ann Arbor: University of Michigan Press, 1976.
Schank, R., and Abelson, R.P. *Scripts, Plans, Goals, and Understanding*. Hillsdale, N.J.: Erlbaum, 1977.
Searle, J.R. (Ed.) *The Philosophy of Language*, Oxford: Oxford University Press, 1971.
Smuts, B. Special relationships between adult male and female olive baboons (*Papio anubis*). Ph.D. diss., Stanford University, 1982.
Strum, S.C. Baboons may be smarter than people. *Animal Kingdom* 88: 12–25, 1985.
Tooby, J. Prospects for an evolutionary psychology. Unpublished manuscript, Harvard University, 1975.
—— The emergence of evolutionary psychology. In *Emerging Syntheses in Science*, D. Pines (Ed.). Santa Fe: Santa Fe Institute, 1985.

——, and Cosmides, L. Evolutionary psychology and the generation of culture, part I. Theoretical considerations. *Ethology and sociobiology* 10: 29–49, 1989.

——, and —— The evolution of revenge. In preparation, a.

——, and —— Evolution and cognition. In preparation, b.

——, and DeVore, I. The reconstruction of hominid behavioral evolution through strategic modeling. In *Primate Models for the Origin of Human Behavior*, W.G. Kinzey (Ed.). New York: SUNY Press, 1987.

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

*Wall Street Journal*. Prestige cards: For big bucks and big egos. April 17, 1985, p. 35.

Wason, P.C. Reasoning. In *New Horizons in Psychology*, B.M. Foss (Ed.). Harmondsworth, U.K.: Penguin, 1966.

—— Realism and rationality in the selection task. In *Thinking and Reasoning: Psychological Approaches*, J. St B.T. Evans (Ed.). London: Routledge and Kegan Paul, 1983.

——, and Johnson-Laird, P.N. *Psychology of Reasoning: Structure and Content*, London: Batsford, 1972.

Wilkinson, G.S. Reciprocal food sharing in the vampire bat. *Nature* 308: 181–184, 1984.

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

Wrangham, R.W. War in evolutionary perspective. In *Emerging Syntheses in Science*, D. Pines (Ed.). Santa Fe: Santa Fe Institute, 1985.