Chapter 5

Human Social Exchange

Evolutionary biology provides a heuristic for guiding psychological theory; the research and theory presented in this dissertation is meant to be an illustration of its potential. This heuristic rests on the recognition that natural selection has produced psychological mechanisms as responses to various 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. Successfully conducted social exchange was a critically important feature of hominid evolution. Natural selection permits the evolution of only certain strategies for engaging social exchange. By studying the nature of these strategies, one can deduce many properties that human 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, reasoning, the organization of memory, and learning. One can also make specific predictions about human peformance on reasoning tests like the Wason selection task.

Chapter 5 examines the nature of the selective pressures on social exchange in human evolution, and what these allow one to infer about the psychological basis for social exchange in

humans. It is divided into three parts which make the following points:

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

5.2: 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.

5.3: 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.

* * *

5.1 Natural selection and social exchange:

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

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

There are laws inherent in the dynamics of natural selection that hold for any species, on any planet, at any time. Many of these laws govern the evolution of social behavior; they constrain the kinds of social behavior that can evolve.

Traits can be thought of as the embodiment of <u>strategies</u> for the propogation of the genes that code for them. By analyzing the dynamics of gene flow through populations, one can determine what kinds of traits will quickly be 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 another in a race to see which one comes to dominate the gene pool. Such games can be mathematically modeled with great precision.* During the last 20 years, game-theoretic models of the dynamics of natural selection have proliferated in evolutionary biology. This process has led to a startling discovery: there are certain strategies that simply cannot be selected for (e.g., Hamilton, 1964; Williams, 1966; Maynard Smith, 1978; Dawkins, 1982). Furthermore, game-theoretic analyses can be used to specify what stategies are likely to be selected for, and what properties these strategies must have. This claim deserves an illustration from the literature of evolutionary biology.

Given an individual, X, define a BENEFIT TO X (B(X)) as any act, entity or state of affairs that increases the number of replicas of a given gene (offspring) which that individual produces through his or her own reproduction. Similarly, define a COST TO X (C(X)) as any act, entity, or state of affairs that decreases the number of gene replicas that individual produces. Just to exhaust the possibilities, let 0(X) refer to any act, entity or state of affairs that has no effect on the number of gene replicas X produces. By so defining the effects which different morphological, physiological, or behavioral traits can have on gene replication through a particular individual, one can compare two alternative traits to see which one leads to greater

* The Modern Synthesis -- the wedding of statistical methods to Mendelian genetics -- brought rigor to evolutionary biology in the 1930's. During the last 20 years that rigor has been substantially enhanced by (1) the identification of the gene as the unit of selection, and (2) the technological ability to create computer models of strategic games.

replication of the genes which underlie it and will therefore spread through the population.

Now, consider this excerpt from Hamilton, 1972:

> A gene is being favored in natural selection if the aggregate of its replicas forms an increasing fraction of the total gene pool. We are going to be concerned with genes supposed to affect the social behavior of their bearers, so let us try to make the argument more vivid by attributing to the genes, temporarily, intelligence and a certain freedom of choice. Imagine that a gene is considering the problem of increasing the number of its replicas and imagine that it can choose between causing purely self-interested behavior by its bearer **A** (leading to more reproduction by **A**) and causing "disinterested" behavior that benefits in some way a relative, **B**. (p. 195)

Hamilton then computes how many replicas of this gene will be produced if it codes for decision rule 1 versus decision rule 2:

For any act, Z, which would benefit **A**'s relative, **B** (Z = B(B)):

Decision Rule 1. If [C(A) of doing Z] > 0, do not do Z.

Decision Rule 2. If [C(A) of doing Z] < [B(B) of receiving Z] <u>discounted by</u> r(A,B) (a fraction denoting the probability that B contains a replica of the gene in question), then do Z.

More replicas of the gene in question will be produced if that gene codes for decision rule 2* rather than decision rule 1. This result holds for <u>every</u> species which can selectively confer benefits on relatives. It is a law inherent in the dynamics of natural selection.

Instead of imagining a gene contemplating various strategies, an entirely equivalent way of considering the same evolutionary problem is to imagine that the two decision rules

---

* gene 2 codes for decision rule 2 which, on average, maximizes an individual's "inclusive fitness" -- his own reproductive success plus his effects on the reproductive success of his relatives, each effect discounted by the appropriate r, the coefficient of relatedness (Hamilton, 1964; Dawkins, 1982).

are embodied in different organisms. One then imagines a tournament pitting Gene 1 (which codes for Decision Rule 1) against Gene 2 (which codes for Decision Rule 2).

In the tournament, two individuals face the same environment. That is, in the first generation both individuals have the same number of relatives, the same number of opportunities/per relative for conferring benefits, and the same set of payoffs associated with particular opportunities. Individual 1 has gene 1, and therefore uses decision rule 1; Individual 2 has gene 2 and therefore uses decision rule 2. Before the tournament starts, genes 1 and 2 exist in equal numbers in the population.

Using this tournament, one can ask: After one generation, how many replicas of gene 1 versus gene 2 exist in the population? How many replicas of each exist after n generations? If one were to run a computer model of this tournament, one would find that after one generation there would be more replicas of gene 2 than gene 1; the magnitude of the difference between them is gene 2's "selective advantage" over gene 1. This magnitude will depend on what payoff and opportunity parameters were specified in the program used. After n generations, where n is a function of the magnitude of gene 2's selective advantage in the tournament's "environment", one would find that gene 2 had "gone to fixation": that gene 1 would represent a vanishingly small fraction of the gene pool, regardless of the absolute size of the population.

Using the same thought experiment one can ask other questions: Once gene 2 has become fixed in a population, is it

vulnerable to invasion by a mutant gene coding for a different decision rule (i.e., is it an Evolutionarily Stable Strategy, an ESS)? If gene 1 is fixed in the population, is it vulnerable to invasion by a mutant gene 2? Will gene 2 sweep the population, or will a stable polymorphism result between genes 1 and 2? Is a gene better off if it codes for a mixed strategy, one that uses decision rule 2 under certain circumstances, and some other decision rule under other circumstances? And so on.

In other words, natural selection theory has a game theoretic structure (Maynard Smith, 1982). This fact can be usefully applied to an analysis of social exchange between unrelated individuals.

In the example above, the decision rules governed a unilateral act; should I, or should I not, benefit my relative by doing act Z? In contrast, social exchange involves two acts: what I do for you (act 1) and what you do for me (act 2). My doing act 1 for you benefits you (B(you)) at some cost to myself (C(me)). Your doing act 2 for me benefits me (B(me)) at some cost to yourself (C(you)). Furthermore, the benefit to you of receiving my act 1 is greater than the cost to you of doing act 2 for me (B(you) > C(you)); likewise, the benefit to me of receiving act 2 from you is greater than the cost to me of doing act 2 for you (B(me) > C(me)). All costs and benefits are measured in inclusive fitness terms: C(X) and B(X) refer to decreases and increases in the inclusive fitness of individual X (see footnote, page 132). If acts 1 and 2 have this cost/benefit structure, we both get a net benefit by exchanging acts 1 and 2. Let's call an interaction that is mutually beneficial,

"cooperation."

At first blush, one might think that natural selection would favor the emergence of psychological mechanisms with decision rules that lead organisms to participate in a social exchange whenever the above conditions hold. After all, participation would result, by definition, in a net increase in the replication of genes underlying a tendency to participate, as compared to genes underlying a tendency to not participate.

But there is a hitch: You can benefit even more by cheating me. If I do act 1 for you, but you do not do act 2 for me, then you benefit more than if we both cooperate. This single fact creates an enormous stumbling block for the evolution of social exchange, a problem that is structurally identical to one of the most famous situations in game theory: the one move Prisoner's Dilemma (e.g., Trivers, 1971; Axelrod & Hamilton, 1981; Axelrod, 1984).*

The Prisoner's Dilemma is a game in which mutual cooperation would benefit both players, but it is in the interest of each player, individually, to defect, cheat, or snitch on the other. It is frequently conceptualized as a situation in which two people who have collaborated in committing a crime are prevented from communicating, while a district attorney offers each individual a lighter sentence if he will snitch on his partner. However, the payoffs can represent anything for which both

* Other models of social exchange are possible, but they will not change the basic conclusion of section 5.1: 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.

players have a similar preference ranking: money, prestige, points in a game, even reproductive success. A possible payoff matrix and the relationship that must exist between variables is shown in Figure 5.1.

Figure 5.1 Payoff Schedule, Prisoner's Dilemma

		you: C	you: D
me	C	me: R = +3 you: R = +3	me: S = -2 you: T = +5
	D	me: T = +5 you: S = -2	me: P = 0 you: P = 0

C = Cooperate
D = Defect
R = Reward for mutual cooperation
T = Temptation to defect
S = Sucker's payoff
P = punishment for mutal defection

T > R > P > S
R > (T+S)/2 *

* For an iterated game, R > (T+S)/2. This is to prevent player's from "cooperating" to maximize their utility by alternately defecting on one another.

Looking at this payoff matrix, one might ask: "What's the dilemma? I will be better off, and so will you, if we both cooperate -- you will surely recognize this and cooperate with me." However, if there is only one move in the game, it is always in the interest of each party to defect (Luce & Raiffa, 1957) -- that is what creates the dilemma.

Let's say 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

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. Let's say the payoff matrix in Figure 5.1 represented dollars: if you cooperate, I get $5 for defecting instead of $3 for cooperating. If you defect, I lose nothing by defecting instead of losing $2 by cooperating.

Figure 5.2 shows that the cost/benefit structure of a social exchange has the same structure as a Prisoner's Dilemma. If I cooperate on our agreement, you get B(you) for defecting, which is greater than the B(you) - C(you) you would get for cooperating. If I defect on our agreement, you get nothing for defecting (this is equivalent to our not interacting at all), which is better than the C(you) loss you would incur by cooperating. The payoffs are in inclusive fitness units -- the numbers listed are included simply to reinforce the analogy with Figure 5.1. In actuality, there is no reason why C(me) must equal C(you) (or B(me) = B(you)); an exchange will have the structure of Prisoner's Dilemma as long as mutual cooperation would produce a net benefit for both of us.

How can a system of mutual cooperation emerge in such a situation? Given an opportunity for exchange, if my decision rule was "Cooperate whenever B(me) > C(me)" and your rule was "Cheat", the genes underlying my decision rule would soon be selected out. For every interaction with a "cheater" I would lose 2 inclusive fitness points, and the cheater would gain 5.

By definition, then, my tendency to cooperate would be selected out, and the "Cheat" decision rule would spread through the population; the number of generations this takes is a function of

Figure 5.2 Social exchange sets up a Prisoner's Dilemma

		you: C	you: D
me	C	me: B(me) - C(me) = +3 you: B(you) - C(you) = +3	me: C(me) = -2 you: B(you) = +5
	D	me: B(me) = +5 you: C(you) = -2	me: 0(me) = 0 you: 0(you) = 0

B(X) = Benefit to X
C(X) = Cost to X
0(X) = X's inclusive fitness is unchanged

how many cheaters versus indiscriminate cooperators are in the initial population (Appendix B shows just how quickly, given some rather generous assumptions). In practice, a population of "cheaters" is a population of individuals who never participate in social exchanges; if you cheat by not doing act 2 for me, and I cheat by not doing act 1 for you, then in fact, we have not interacted at all -- we have had no effect on one another.

You might object that real life is not like a Prisoner's Dilemma, because real life exchanges are simultaneous, face-to-face interactions. You can directly see whether I am about to cheat you or not. If I show up without the item I promised, then you simply do not give me what I want. This is certainly true in a 20th century market economy, where money is used as a medium of

exchange. But in nature, most exchanges are not, and cannot be, simultaneous. For example:

1. A common "item" of exchange between primates is protection from conspecifics and predators. Two or more individuals develop coalitional relationships for mutual defense, aggression, or protection (e.g., baboons: Hall & DeVore, 1965; chimps: Wrangham, in press; de Waal, 1982). If you are attacked, and I come to your defense, there is nothing you can do, at that time, to repay me. My repayment will come when I am attacked and you come to my defense (I hope!).

2. We are foraging for patchy resources. You find a tree laden with more fruit than you can eat by yourself; you give a shout to guide me to it. There is nothing I can do to repay you on the spot. Your repayment will come in the future when I let you know about a similar find -- you hope (e.g., birds: Ward & Zahavi, 1973; bats: McCracken & Bradbury, 1981; chimps: Goodall, 1968, 1971).

3. In cooperative hunting, there is only one kill at a time, and usually only one or two individuals actually make the kill. Those who actually make the kill claim the most, but they share the rest of the meat with the others on the hunt, trusting that they will share one of their kills at some future time. Again, repayment on the spot is impossible.

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 in exchange for sexual access when the females wean their infants and become fertile again (e.g., Smuts, 1982; Strum, 1985). His repayment, by necessity, comes at a much later time.

The opportunity for on-the-spot repayment 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 continually shifting: 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. Frequently, simultaneous needs or windfalls cannot be turned into opportunities for mutual aid: if two individuals are attacked simultaneously, neither is free to help the other; if they find two food sources simultaneously, neither benefits from the other's windfall.

Thus, in the absence of a widely accepted medium of exchange,*

most exchanges do constitute a Prisoner's Dilemma. You must decide whether to benefit me or not without any guarantee that I will return the favor in the future. This is why Trivers (1971) describes social exchange in nature as "reciprocal altruism." I behave "altruistically" (i.e., I incur a cost in order to benefit you) at one point in time, and you reciprocate my altruisic act in the future. If you do, in fact, reciprocate, then our "reciprocally altrusitic" interaction is properly described as an instance of delayed mutual benefit: neither of us has incurred a net cost, both of us have gained a net benefit. Obviously, however, if only one interaction is involved -- that is, if we are playing a Prisoner's Dilemma game with only one move -- I would be a fool to reciprocate your altruistic act, and you, knowing this, would be a fool to do it in the first place. So we are back to square 1: mutual defection is in both of our interests.

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,**

---

* Indeed, such factors are exactly why it is so useful to have a medium of exchange. I don't have to be able to provide the particular goods or services you want because you can convert money from me into anything. Furthermore, money permits a simultaneous exchange, in which I can, in fact, withold my money if I see that you intend to cheat me, and vice versa.

** 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 & Raiffa, 1957).

and 3) they do not value later payoffs by too much less than earlier payoffs (Axelrod & Hamilton, 1981; Axelrod, 1984). If you and I are making a series of moves rather than just one, your behavior on one move can influence my behavior on future moves. If you defect when I cooperated, I can retaliate by defecting on the next move;* if you cooperate when I have, I can reward you by continuing to cooperate. In an iterated Prisoner's Dilemma, a system can emerge that has incentives for cooperation and disincentives for defection.

For example, cooperation can be selected for if it is governed by a decision rule that says: "Cooperate with individuals who have cooperated with me in the past; defect with individuals who have a history of defection." Using the payoff matrix in Figure 5.2, it is clear that a strategy like this could be selected over an "always cheat" strategy. The mutual cooperators would get strings of +3 inclusive fitness points, peppered with a few -2s from a first trial with a cheater (after which the cooperator ceases to cooperate with that individual). In contrast, mutual defectors would get strings of zeros,

---

* 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 <u>reliably</u>. 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. Therefore, it is by no means clear that the use of violence is the most cost efficient way to foster cooperation, especially in a one move game. 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.

peppered with a few +5s from an occasional first trial with a cooperator (after which the cooperator never cooperates with that individual again).

A number of strategies permitting selective cooperation are possible, but one that has been particularly successful in recent investigations is called TIT FOR TAT (Axelrod & Hamilton, 1981; Axelrod, 1984). It is a very simple strategy in which: 1) I cooperate on the first move, and 2) I do whatever you did on the previous move. If you cooperate on move 1, then I cooperate on move 2; if you defect on move 1, then I defect on move 2. TIT FOR TAT can be used to illustrate the selective advantage of selective cooperation.

Table 5.1 is designed to give you an idea of how a TIT FOR TAT decision rule stacks up against an "always cheat" decision rule (CHEAT) and a mixed strategy rule (MIXED) in a round robin tournament. The mixed strategy rule is a TIT FOR TAT program that slips in some cheating on the side. After a mutually cooperative move, it tries to rack up points by defecting. If it succeeds in earning T, it immediately "apologizes" for its defection by cooperating on the next move, in an attempt to restore mutual cooperation. If MIXED does not succeed in earning T (i.e., if its partner also defected), it "retaliates" by defecting on the next move.

As you can see, TIT FOR TAT earns more points in this round robin than either MIXED or CHEAT. Because points stand for replicas of genes coding for each decision rule, this means that TIT FOR TAT genes would spread through the population, eventually displacing MIXED and CHEAT. This result is not an artifact of

Table 5.1 Round robin tournament pitting TIT FOR TAT (TFT) versus CHEAT versus MIXED

	TFT		v. MIXED		SUB TOTALS	TFT		v. TFT		MIXED		v. MIXED	
1	C	+3	C	+3	3,3	C	+3	C	+3	C	+3	C	+3
2	C	-2	D	+5	1,8	C	+3	C	+3	D	0	D	0
3	D	+5	C	-2	6,6	C	+3	C	+3	D	0	D	0
4	C	+3	C	+3	9,9	C	+3	C	+3	D	0	D	0
5	C	-2	D	+5	7,14	C	+3	C	+3	D	0	D	0
6	D	+5	C	-2	12,12	C	+3	C	+3	D	0	D	0
7	C	+3	C	+3	15,15	C	+3	C	+3	D	0	D	0
8	C	-2	D	+5	13,20	C	+3	C	+3	D	0	D	0
9	D	+5	C	-2	18,18	C	+3	C	+3	D	0	D	0
10	C	+3	C	+3	21,21	C	+3	C	+3	D	0	D	0
		+21		+21			+30		+30		+3		+3

	TFT		v. CHEAT		CHEAT		v. CHEAT		MIXED		v. CHEAT	
1	C	-2	D	+5	D	0	D	0	C	-2	D	+5
2	D	0	D	0	D	0	D	0	D	0	D	0
3	D	0	D	0	D	0	D	0	D	0	D	0
4	D	0	D	0	D	0	D	0	D	0	D	0
5	D	0	D	0	D	0	D	0	D	0	D	0
6	D	0	D	0	D	0	D	0	D	0	D	0
7	D	0	D	0	D	0	D	0	D	0	D	0
8	D	0	D	0	D	0	D	0	D	0	D	0
9	D	0	D	0	D	0	D	0	D	0	D	0
10	D	0	D	0	D	0	D	0	D	0	D	0
		-2		+5		0		0		-2		+5

CONTENDER'S SCORES:

		opponents						
		MIXED		TFT		CHEAT		
contenders	TIT FOR TAT:	21	+	30	+	-2	=	49
	CHEAT:	5	+	5	+	0	=	10
	MIXED:	3	+	21	+	-2	=	22

the particular strategies it was pitted against in Table 5.1.

Robert Axelrod conducted a round robin computer tournament

in which TIT FOR TAT was pitted against 62 other entries. All entries were submitted by sophisticated students of the Prisoner's Dilemma, including professors of psychology, biology, and political science. TIT FOR TAT achieved the highest average score (Axelrod, 1984). Its success appears to be due to four factors:

1. TIT FOR TAT is "nice": it never defects first
2. When its opponent defects, TIT FOR TAT retaliates; hence TIT FOR TAT is not exploitable
3. TIT FOR TAT is "forgiving": if its opponent initiates cooperation after having defected, TIT FOR TAT cooperates on the next move; it does not get caught in endless chains of recriminations (as in MIXED v. MIXED)
4. TIT FOR TAT is so clear and consistent that, once encountered, it is easily recognized, and its non-exploitability is easily appreciated.

The authors of all entries submitted knew that TIT FOR TAT had won a previous tournament of 12 entries. Furthermore, they had been given an extensive analysis of the properties that had led to its success. Some authors submitted mixed strategies that usually played TIT FOR TAT, but tried to get away with occasional cheating. Others, guessing that the analysis provided would prompt many authors to submit "nice" strategies, submitted "exploitative" strategies -- strategies designed to take advantage of "nice" entries. In general, the exploitative strategies won an occasional battle but lost the war, earning the lowest average scores in the round robin.

Axelrod also conducted a simulation of natural selection over time: The more points a strategy earned in one "generation" (round robin), the more "copies" of that strategy competed in the next "generation." Over the generations, TIT FOR TAT and other

nice-but-retaliatory strategies came to dominate the population. The exploitative and mixed strategies eventually went "extinct." TIT FOR TAT always had the largest share of the "gene pool": by the 1000th and last generation, its representation in the gene pool was still growing at a faster rate than that of any other strategy.

Other calculations demonstrated that a very small cluster of TIT FOR TATTERs can invade a population of cheaters, even if very few of their interactions are with each other. Futhermore, it can be mathematically demonstrated that TIT FOR TAT is an Evolutionarily Stable Strategy (ESS): no "mutant" strategy can invade a population composed primarily of TIT FOR TATTERS, either singly or in small clusters. The average performance of a TIT FOR TATTER in a population of its fellows is higher than the average performance of any possible newcomer.

The details of TIT FOR TAT are not what is important about this story. The key point, which TIT FOR TAT illustrates, is that a cooperative strategy can invade a population of non-cooperators if, and only if, it cooperates with other cooperators and cheats on cheaters. Indiscriminate cooperation cannot be selected for in any species. We 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.

5.2 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 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.* 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 adaptiveness 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. Moveover, in all probability these individuals, like modern hunter-gatherers, were closely related; kin selection can be a tremendous aid to the evolution of cooperation (Trivers, 1971; Axelrod & Hamilton, 1981).

The intellectual capacities of early 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

* 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-R)/(T-P) and (T-R)/(R-S) (Axelrod, 1984).

achieve a goal. The exploitation of a new savannah and woodland niche -- made possible by tool use -- allowed individuals to acquire food items too large to be consumed by a single individual (Tooby & DeVore, 1985).* This created the perfect opportunity to provide a large benefit to another individual at a very 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 indicates that Pleistocene hunter-gatherers, like their modern counterparts, engaged in extensive food-sharing (e.g., Issac, 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** created situations in which coordinated, cooperative behavior could produce vast payoffs. Perhaps one of the best examples are the "profits" to be made through cooperative hunting. Acting together, several armed men can bring down a wooly mammoth; acting alone, a single armed man cannot.

These conditions set the stage for the coevolution of a tightly interwoven complex of adaptations that made cooperation

* And which could not be stored for later use without spoiling -- early hominids had no refrigeration!

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

more and more profitable (Tooby & DeVore, 1985). Cooperative hunting provided a compact and nutritious food source that provided an efficient means for males to invest in offspring; leading to mechanisms to insure their paternity; leading to (1) more closely related subsets of individuals within bands, creating larger payoffs for cooperative behaviors and more group stability (which creates even more opportunities for cooperation), and (2) even greater payoffs for male parental investment in offspring; leading to more male parental investment; which allows larger brains and longer periods for maturation and learning; leading to more efficient cooperation and tool use, and therefore to even more nutritious food sources from both hunting and gathering; making it more efficient to devote metabolic resources to brain over brawn...and so on, each condition circling back to amplify the effects of the ones before it, until today cooperation for mutual benefit is a pervasive and inextricable aspect of all human cultures.

Reconstruction of the exact causal chain that led to the evolution of cooperation is still a matter of debate (cf. Kinzey, 1985). 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.

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 more "item-specific" the algorithms regulating exchange can (and should) be: What counts

as "error" -- cheating or under-reciprocating -- can be more closely defined, increasing the accuracy of one's mental accounting system and the accuracy of reference (see section 5.3). In contrast, human algorithms for regulating social exchange should be able to handle a wide and ever-changing array of "items" for exchange: tools, 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 they require to function properly -- will have some human-specific properties. These will be discussed in the next section.

5.3. 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 & 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 & 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. 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. A computational theory of social exchange must be powerful enough to (1) permit the realization of a "possible" social exchange strategy, that is, a strategy that can be selected for, and (2) exclude "impossible" strategies, that is, strategies that cannot be selected for.

The ability to engage in a possible strategy of social exchange presupposes the ability to solve a number of information-processing problems. The problems most specific to social exchange will be incorporated into a "grammar of social contracts" in the second half of this section. A grammar of social contacts is the set of assumptions about the rules governing a particular social exchange that must somehow be incarnated in the psychological mechanisms of both participants. It is the aspect of the computational theory of social exchange most relevant for understanding performance on the Wason selection task.

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, like mating, pair-bonding, parenting, and aggression. Before progressing to the grammar of social contracts and its implications for performance on the Wason selection task, 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 items one perceives as causally connected to biologically significant variables as costs and benefits; 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 slowly and clumsily, by domain general mechanisms like 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 Fodor's or Marr's terminology, **mental organs** or **cognitive competences** in Chomsky's 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. I shall briefly sketch a few of these implications, occasionally citing relevant data.

5.3.1 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 & 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. But cooperation can evolve only if it is based on reciprocation. 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.* One need not rely on "preliminary hunches" (Carey & 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 evolutionary theory.

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, Bahrick & Wittlinger, 1975). Patients with a lesion in a specific part of the right hemisphere develop a selective deficit in their ability to recognize faces,

---

* Organisms that lack the ability to recognize different individuals can also evolve a limited ability to cooperate, but only by restricting their interactions to a very few partners with whom they are in constant and/or exclusive physical proximity (Axelrod & Hamilton, 1981).

called prosopagnosia (Gardner, 1974). Carey & 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 individual human gaits (Cutting, Proffitt, & Kozlowski, 1978; Kozlowski & 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 previous interactant, and not, for example, a stranger, a mate, or an offspring. Second, once an individual has been identified as a previous interactant, information regarding 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 section 5.1, most 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.* There is no need to remember how much you owe or are owed, because there is no owing: each transaction is either a complete exchange or a complete defection. The potential for cheating is much higher, however, in exchanges in which reciprocation is

* 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.

delayed; once you have conferred a benefit, you cannot take it back. To be able to "call in your markers", you must be able to keep track of who owes what. Consequently, the capacity for engaging in transactions in which reciprocation is delayed requires a mental accounting system for keeping track of who owes who what (note: Proposition 5 also applies to this accounting system).

The extent of the history of interaction that must become available to the decision procedure that regulates whether you agree to participate in a particular social exchange (and whether any of these facts need be consciously recalled) will depend on the details of the particular decision procedure humans have evolved. TIT FOR TAT requires only that the last transaction with each interactant be recalled. But TIT FOR TAT operates 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. 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. 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 behaved in those transactions, (3) the size of payoffs to

both parties in previous transactions, (4) whether his tendency to cheat varied with the size of the payoff involved, (5) whether the conditions governing his tendency to cheat have been shifting over time, (6) his (relative) aggressive formidability, (7) how likely one is to meet that individual in the future (e.g., one of you is moving away or likely to die soon), and (8) whether one of you accepted a past benefit but has not reciprocated yet.

A decision procedure that used such data, current behavioral cues,* and the payoff matrix for the current interaction to compute the conditional probability that one's partner will cooperate, might be better adapted to the complexities of exchange in nature.** 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, and activated quickly and effortlessly when an opportunity for exchange with that person arises. When the payoff matrix of the current

---

* For example, my facial expression might tip off my intention to cheat you. 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 & Diamond, 1980).

** 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 which became more likely to cheat in later interactions. In a sense, it failed because it assumed that competitor programs had static "personalities."

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 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 ruin 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* affective mechanisms that communicate cost/benefit information: they may cause you to cry, turn your back on me, scream at me, hit me. The extent and nature of the overt aspects of your affective reaction communicates to me your view of the extent of my wrong doing: 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. Emotion communication can be viewed as one way individuals communicate cost, benefits, and behavioral

---

* I say "other" because I see no principled way of drawing a dividing line between emotion and cognition. The flood of memories you experience when I betray you is as much a part of your "emotional reaction" as your turning red and punching me out (see Tooby, in press; Tooby & Cosmides, in preparation). 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, non-linguistic organisms can also engage in social exchange -- however, the range of items they can exchange is necessarily more limited. For example, chimps recruit support from others in aggressive encounters, and frequently form long-term coalitional relationships. These coalitions are social exchanges in which the exchanged "item" is mutual aid in fights. A chimp 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 chimp 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 chimp 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 chimps in the Arnhem chimp 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 chimp]. 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 refering 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 nice pears? Do I want you to hold a branch-ladder so I can climb into a tree which 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 & 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 kidnapping 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 kidnapped 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 abducted into a hamadryas male's harem, 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 it 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 non-linguistic 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). I 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 which 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 and individuals who cannot yet speak a language (small children).

2. Emotion signals can function like intersubjective rulers, 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 trembles 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 rulers, 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 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, or publicly posting a long list stating one's preference hierarchy...with periodic updates!*

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 fight for fight, fight for sex, sex for sex, food for food, fight, or sex, groom for groom, groom for fight, food, or sex...and not too much else. 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.** But a natural language permits a potentially infinite number of arbitrary, discriminable symbols to be

---

* Actually, a Santa's 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?

** 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 + model system amounts to a referential semantics (Gleitman & Wanner, 1982).

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.

This property of language opens the vast realm of human adaptations associated with planning and tool-use to social exchange. Tool technology continually changes,* with new tools being invented constantly. New technologies enable new and constantly changing opportunities for coordinated, cooperative behaviors which 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 non-linguistic gestural system -- routines like the chimps' ladder expedition, that are discovered quite publicly in the context of an emotionally salient event,** and don't require long periods of planning.

---

* At least for Homo sapiens sapiens. Homo erectus' tool kit stayed identical over a wide range of different environments -- from Asia to Africa -- for over 1.5 million years (Pilbeam, personal communication). 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.

** The Arnhem chimps discovered the ladder trick when one screaming chimp, fleeing from a very public attack, bounded up a broken branch that happened to be resting against a tree.

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 so limited, each item could, in theory, have a unique cost/benefit weighting associated with it that is shared by most other members of the species (e.g., ten grooms deserves one fight, a season of protection by a male deserves exclusive sexual access at the height of estrus, etc.). In other words, each item could have a preprogrammed, universally acknowledged, "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* or intentionally -- via emotion signals. Indeed, there is rudimentary evidence suggesting that some aspects of the

* 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 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.

acoustic expression of emotion in humans have been integrated into our species-specific language capacity in ways that facilitate the communcation of values and intentions (Cosmides, 1983).

Proposition 4. One must be able to model the values of other individuals.

In some ways, Propøsition 4 is just the flip 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 suggest 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 configurational changes in tall grass as it is blown by the wind or the sound of a motor starting up.* Attention should be more sustained for emotion signals emitted by a potential interactant

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

-- 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, p. 33) 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 I am hungry, I moan because the thing I value -- food -- is <u>not</u> present. You must infer my desire for food from my moan, even though there is no spatio-temporally contiguous event in which the signal (my moan) and the referent (food) are both present.

Evolutionary theory provides a rich heuristic base for developing theories about what kinds of preference information is included in our referential 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, political structure, 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 -- a few of these are listed in Box 5.1. In addition to being very incomplete, this list is extremely simple, in that it assumes

"all else equal". There are complex interactions among these factors that evolutionary theory speaks to. Hence, a "cognitive list" is not enough: the algorithms that guide our marketing research should include cost-benefit analysis procedures that allow one to take these complexities into account in modeling other people's values.

Although researchers from Bartlett (1932) to Schank & 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 theory as a heuristic rudder, 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.*

---

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

Box 5.1 Typical Human Preferences

The following is a very minimal list of events and states of affairs that the typical human is likely to prefer, all else equal; I have made no effort to distinguish fundamental goals from behaviors or traits reliably paired with fundamental goals. These preferences are suggested by evolutionary theory; however, only psychological research can establish which ones have been incorporated into the human motivational system.

OFFSPRING: having offspring over barrenness, low child mortality, having as many offspring as available investment will allow, own over other's offspring, kin's offspring in proportion to degree of relatedness, kin's over unrelated offspring, fertile children, chastity of daughters when males control means of investment, sons good at acquiring resources, resources distributed equally to grandchildren (different own children's preference), inhibition from harming own, kin's, and friends' offspring (in that order); MATING: outbreeding over incest, sex over celibacy, a pair-bonded mate; FEMALE MATING PREFERENCES: a male who can invest in her offspring over a male who cannot, investment fidelity in a mate, sexual fidelity (especially insofar as it is related to investment fidelity), a mate who is also willing to invest in her kin, willingness to forgo a male's appearance if he is a good investor, the ability to live near female's male kin, being sole wife, being first wife over being co-wife, investment directed at one's own offspring rather than husband's offspring by co-wife, having a sister as co-wife over stranger as co-wife, marrying when she is young, having lovers who invest (as long as husband doesn't find out), affection more than sex; MALE MATING PREFERENCES: paternity certainty, sexual fidelity in wife, females whose appearance is characterized by cues suggesting high reproductive value and/or fertility, opportunities for sex with other females, marrying women at peak of their reproductive value, sex out of wedlock with women at peak of their fertility (female fertility peak being somewhat later than peak reproductive value), sex more than affection (except with post-reproductive wife), having as many wives as ability to invest will bear; FOOD AND OTHER INVESTMENT: food over starvation, for kin over friends or strangers, for friends over strangers, for oneself in preference to one's sib (up to point determined by degree of relatedness), willingness to protect offspring, other kin and friends from predators (in that order); SOCIALITY: having cooperative relationships (friends), reciprocation over nonreciprocation, aiding a friend over aiding a stranger, cheating when it will remain undetected, not being ostracized from one's social group, own death over death of all possible offspring, willingness to commit infanticide if keeping the child will result in the loss of older child, willingness to kill one twin if keeping both will result in loss of both, power over powerlessness, (for males) having powerful coalitional allies, brothers over friends as allies, being aggressively formidable, going to war when the probability of achieving a net gain in captured women and resources is sufficiently high, going to war when wife, children, and resources are threatened by other males; HEALTH: health over injury or disease, not having diseased persons or their effluvia nearby, avoidance of disease-breeding filth, avoidance of decomposing bodies, fresh food over rotting food, avoidance of poisonous animals (spiders, snakes, etc.), avoidance of predators.

Veridical recall of stories that violate the assumptions about human preference instantiated in our motivation scripts should be difficult. 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 questions 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 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.*

---

* 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 <u>especially</u> interesting -- Nightly News coverage of a fire at your neighbor's house verus a fire in Charlestown; a fist fight in the halls of William James versus a fist fight in Southie.

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 especially boring.

The learning mechanisms that guide our 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 be filed under his or her "name." When the opportunity to acquire more preference information about an individual arises, the model appropriate to that individual must be easily retrieved, not just any person-model, or a model of average preference. "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.* Learning

> "Smith values W more than X more than Y more than Z", and
> "Jones values Z more than X more than Y more than W"

is useless unless it increases your ability to make offers that maximally 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 too 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

* 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.

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 fishes 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, some items should be more readily perceived as costs and benefits -- those for which the perceiver can ascertain a clear causal link to biologically significant variables like offspring, kin, sex, food, safety, shelter, protection, aggressive formidability, and dominance. For example, a Mr. Michael Pastore of Dallas recently made the following comment in an interview with 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." ("Prestige cards: For big bucks and big egos." 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.* Knowing this, one should readily assume that Mr. Pastore perceives owning an American Express Platinum Card as a benefit, and that if he did not own one he would probably be willing to give up other items in order to acquire one. It is a suitable item for social exchange.

5.3.2 The grammar of social contracts

A grammar of social contracts specifies the properties that must be embodied by a Darwinian algorithm for reasoning about social exchange. It incorporates the stategic constraints outlined in 5.1 and the ecological constraints outlined in 5.2

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, because all the necessary specifications are not spelled out

* 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, in press).

directly in speech. This grammar creates the "cohesion of discourse" (Wason & Johnson-Laird, 1972, p. 92), and the cohesion of behavior, in interactions involving uncoerced exchange. It constitutes 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).

Unlike the exchange algorithms of cleaner fishes 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 represent-ations (see 5.2; Proposition 5). The proposed grammar of social contracts is therefore expressed largely in cost/benefit terminology.

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 5.1). 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.

* * *

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 approve my thesis, then I'll give you a million dollars." (Thought I'd get your attention. It's not a sincere offer though -- see the notion of a "sincere offer" below.) P stands for "you approve my thesis" and Q stands for "I'll give you a million dollars." Likewise, not-P stands for "you do not approve my thesis" and not-Q stands for "I do not give you a million dollars".

At the time of my 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. Call this baseline 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 does have something 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 that is irrelevant for this analysis.

What conditions must hold for you to accept my offer?

Let us consider what conditions must hold for you to accept my 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 my offer, one of those expectations about the future must be not-Q -- you do not expect to be receiving $1m from me. If not-Q comes to pass, your utility level will not have moved from your zero level baseline, 0(you).

Q -- receiving $1m 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.* Let's say you value having a million dollars (Q) more than you value not having a million dollars (not-Q); with a million dollars you could feed the starving masses, sail a yacht to Tahiti, whatever. Then Q -- having a million dollars -- constitutes a benefit to you. You will not accept my 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.** In my offer, P -- approving my thesis -- 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 approve my thesis anyway it would be silly of me to offer you the million dollars).*** If P is <u>not</u> something you expected to do in the absence of my offer,

---

* Presumably there are costs and benefits associated with any action. More precisely, B(you) is a net benefit -- the benefits to you of receiving $1m are greater than the costs to you of receiving $1m.

** Again, this is a net cost -- the cost to you of approving my thesis is greater than the benefit to you of approving my thesis.

*** P does not have to be a C(you) for you to accept my contract, although I must <u>believe</u> that it is a C(you) in order to offer the contract in the first place. You could be trying to "snooker" me into offering this contract by dissembling about your real intentions. Perhaps you have been planning to approve my thesis all along, but led me to believe that you are not planning to approve it so I would make you an offer. See below: "Snookering"

then, in your value system, not-P (not approving my thesis) 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 -- you will be no worse off than if my offer had never been made. Let's say that my thesis is terrible and approving such a work would violate your ethical standards, cause you to risk losing your tenure, be the first step in the downfall of Western civilization...whatever. Then P -- approving my thesis -- decreases your utility and is therefore a cost to you, C(you).

Stated in terms of <u>your</u> 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. We will call B(you) minus C(you) your "profit margin". For you to accept my offer, a million dollars must be more of a benefit to you than approving a terrible thesis is a cost. If this is not the case there would be no point in your entering into the contract; it would not increase in your utility. The greater the magnitude of B minus C (the greater your profit margin), the more attractive the contract will appear (an offer of one million dollars for approval of my thesis is more attractive than an offer of one thousand dollars). A contract that reversed this constraint (such that C>>B) sounds perverse. For example, I doubt anyone would be silly enough to make the offer: "If you break your arm then I'll give you a penny." In fact, 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.

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, in this case, me. What conditions must hold for me to be willing to offer a contract? First, I must believe that not-P (your not approving my thesis) 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 thesis approved must increase my utility, it must be a BENEFIT TO ME (B(me)). Third, not-Q -- not giving you $1m -- 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 $1m, and if not-Q comes to pass, I will not have moved from my zero utility baseline.* Fourth, if not-Q is part of my zero baseline, then Q -- giving you $1m -- 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 getting my thesis approved is worth more to me than relinquishing the million dollars).

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 approve my thesis and I am willing to

* 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 $1m regardless, but want to get as much as I can in return. See below: "Snookering."

give up $1m for that approval; but I am not willing to give up my $1m without your approving my thesis.)

In your value system, "If P then Q" translates to:

"If C(you) then B(you)."

("If I (thesis reader) incur the cost of approving Leda's thesis, then I will get the benefit of receiving $1m from Leda".)

However, in my value system the same offer translates to:

"If B(me) then C(me)."

("If I (Leda) get the benefit of your approving my thesis, then I will incur the cost of relinquishing my $1m to you.")

As you can see, P represents a different utility level to me than it does to you. Ditto for Q. In a well-formed social contract -- 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 5.2.

Table 5.2 Cost/Benefit translation of my offer into your value system and mine.

My offer: "If P then Q"
("If you approve my thesis then I'll give you $1m")

		Your point of view	My point of view
P	(you approve my thesis)	C(you)	B(me)
not-P	(you do not approve my thesis)	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)

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),

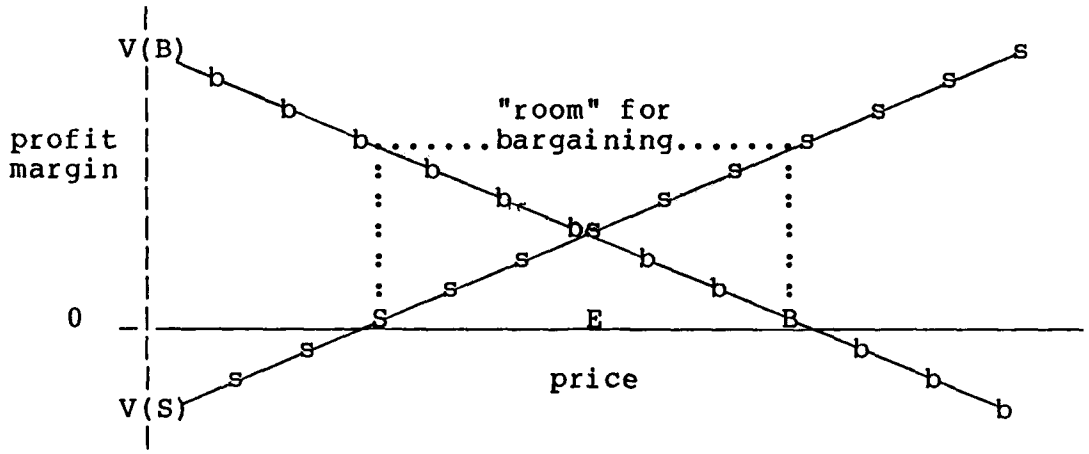
P (approving my thesis) might be a smaller cost to you than Q (giving up $1m to you) is to me (or vice versa). Likewise, Q might be a larger benefit to you than P is to me. These assymetries may lead to a difference in the magnitude of our profit margins (B minus C). Unequal profit margins invite bargaining: you attempt to increase your "profit margin" by paring down mine, and vice versa. Bargaining results in a zero sum game as long as both our profit margins are positive because more B(you) per unit C(you) corresponds to more C(me) per unit B(me). (See Figure 5.3; for a fuller account of these bargaining 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 of us, we have both benefited from the exchange. For this reason, I 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" should be reserved for the violation of a contract.

Snookering

> 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 3 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 woman in the joke has been "snookered."

Figure 5.3 Haggling over the price of a used car. Adapted from Tooby, 1975*



* V(B) represents the value of the used 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 used 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 curve to B. The price range between S and B (the shaded zone) 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 radical 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 <= 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) squawks that labor unions are strangling the company when workers succeed in pushing the price of labor up such that E < price <= 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 profit at any price between S and B.

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

Man's offer: "If you sleep with me 3 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)"

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, 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. Above, we saw that when a contact is offered:

1. not-Q (not giving the woman $15,000) is <u>usually</u> part of the zero utility baseline of the person offering the contract, and
2. Q is therefore a cost to the offerer (Q = C(man)),

but that these are not <u>necessary</u> conditions for making an offer. In the joke, Q was part of the man's zero level baseline: he had planned to give the woman the $15,000 all along. However, it is

a necessary condition of the woman's acceptance that she <u>believe</u> that, in the absence of the offer, not-Q would come to pass. If she expects that Q will happen regardless (Q = 0(woman)), then her utility is decreased by accepting the contract: it is decreased by the magnitude of C(woman).

For any proffered contract of the form: "If you do P then I'll do Q", the acceptor has been "snookered" 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 snookered 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 snookering him by getting the added benefit of $15,000 (until she heard the punch line!). 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. Snookering is different from cheating: In snookering both parties have, technically, honored their contractual obligations. This is not the case with cheating.

Summary so far

The conditions that hold when an individual sincerely offers or sincerely accepts a social contract are shown in Table 5.3. 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 acceptor'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

Table 5.3

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)"	

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 5.4 shows what conditions hold when one person offers or accepts a contract sincerely, but the other person snookers the sincere person. The sincere person believes the contract fits the conditions specified in Table 5.3. However, the snookerer believes the contract fits the criteria specified in Table 5.4. Furthermore, if the sincere person were to find out that she had been snookered, she would share the snookerer's view of the contract's cost/benefit structure.

Table 5.4

SNOOKERING:
Cost/Benefit relations when a sincere party makes a social contract with a snookerer.

	My offer: "If you give me P then I'll give you Q."			
	I try to snooker you; You accept sincerely		You try to snooker me, I offer sincerely	
	If you knew what I knew, we would both believe:		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)"	

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.

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, p.369-370, on Grice).

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

1. I want you to give me P,

2. My offer fulfills the cost/benefit requirements of a sincere contract (listed in Table 5.3),*

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 on the behavior of another person. Some philosophical niceties are discussed in Box 5.2 -- however, these points are not essential to the rest of the chapter.

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 <u>accepts</u> the contract. In contrast, the conditions for a promise hold regardless of whether the hearer consents. Making a promise is a unilateral act; making a contract is not.

---

* An offer that, by virtue of its propositional content, is clearly an insincere contract might be considered snide, or a veiled insult. "I'll give you a dollar if you sleep with your mother" is an insult casting aspersions on your character, which has been thinly disguised as a contractual offer.

Box 5.2 Some Philosophical Niceties (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, snookerers mean their offer 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.

In saying that you accept my offer, you mean that you understand, and agree to comply with, the conditions specified in 1-9 (above). It is like saying "roger wilco": Transmission received (roger), will comply (wilco).

At first blush it might seem that a contract actually expresses a biconditional "Q if and only if P", and will therefore have the same truth table (see Chapter 1 for the truth table of a biconditional). 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 actually a biconditional 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 spoils), 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. Evolutionary theorists have not yet investigated what conditions determine the optimal size of X. I 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 chimps discussed in

Proposition 3, Luit allowed Puist to punish him for his defection. I say "allowed" because Luit is far stronger than Puist, and in a pure test of strength Puist would not have a chance against Luit. To do otherwise would have signaled a drastic change in 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 closely 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)) I am entitled to (however, you must offer it). Failure to accept a benefit one is entitled to may be foolish (and rare -- such behavior would have been strongly selected against), but it does not violate the terms of the contract.

Looking for cheaters

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,* a private agreement or a law of your social group.

Indiscriminate cooperation cannot be selected for in any species. The game-theoretic structure of natural selection theory dictates that social exchange can evolve only if it is governed by a strategy that demands reciprocation. We must cooperate with cooperators and cheat on cheaters. This means our social contract algorithms must include procedures that allow us to quickly infer whether someone has cheated -- or intends to cheat -- on a social contract.

Let's say 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

* 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 US contract law -- a contract is invalid unless some "consideration" has changed hands -- even a symbolic $1 will suffice.

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 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 5.5.

Table 5.5 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)	

As mentioned in Proposition 5, social contract algorithms in humans should be item-independent; they should represent items of exchange as costs and benefits to the participants, and operate on those representations. One cannot look out for cheating unless one can model 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.

This means that for any given exchange, two descriptions of

each item must be computed by the social contact 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 5.5). 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.* There is an analogy here with the grammar of a language. 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 5.3 (or 5.4, if one person is snookering). One would expect these cost/benefit structures to be the descriptions from which participants construct paraphrases and reconstruct the course of the interaction from memory.

Inference procedures for catching cheaters should operate on a cost/benefit description of the contract from the potential cheater's point of view. These procedures should allow one to quickly 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)
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)

---

* Although one might predict that an offer phrased in terms of the potential acceptor's value system might sound more attractive, indicating that the offerer <u>really</u> understands (has a good model of) what the potential acceptor wants!

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 she fulfills her obligation by paying C(X). One keeps an eye on X in situation 4 to make sure she does not illicitly abscond with B(X), to which she is not entitled.

These "look for cheaters" procedures provide the key to understanding performance on the Wason selection task when its propositional content instantiates a social contract. This will be empirically demonstrated in the next chapter.

* * *

I doubt that most people would guess that the structure of a simple, straightforward social exchange is as complex as this chapter shows it to be. But then, that 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.

*

At the beginning of this section I claimed that the grammar of social contracts can be expected to regulate how we think about social exchange; I now feel obligated to provide at least one ecologically valid example from the tool-using hominids of late 20th century America. In reality, are people concerned with

reciprocation and avoiding cheaters? Enjoy the article quoted in Box 5.3 -- I think you will find that the framework proposed in this chapter makes "The Cracker's" reasoning perfectly comprehensible!

Box 5.3 Exchange of "tools" in the Computer Age

The following is excerpted from an article in Popular Computing by a computer hacker named Bill Landreth, alias "The Cracker". He is particularly skilled at acquiring new "accounts", that is, at cracking the access codes of large corporate computer systems. He explains the cost/benefit factors governing his willingness to exchange information about the tools of his trade...

> Information is the currency of the hacker's bulletin-board culture, and trading is the means of exchange. Accounts take a lot of work to get, so most hackers are unlikely to post information publicly when they can trade it for more information from other hackers. In addition, an average hacker acquires only four to five new accounts in a year, and all but maybe one of these accounts die within six or seven months. That same hacker could, however, trade those four or five accounts four or five times each, and those exchanges would net him as many as 25 different accounts in a year.
>
> ...I posted messages on hacker bulletin boards, advertising that I was willing to trade any information I had. I realized that I could be accepted as a bona fide hacker relatively quickly by trading only the highest-quality information. Within a few months of my first postings, the word began to get around: The Cracker is OK.

And on "cheaters"...

> A more important reason for trading, though, is to keep information out of the hands of novices. Often, when novices get hold of publicly posted information, they abuse it by sending obscenities to the system operator, destroying information, changing passwords, or removing accounts. Moral arguments aside, hackers dislike this kind of abuse because accounts that are abused are discovered and die quickly. (p. 64)
>
> ...it became very difficult to tell who you could safely trade information with. Sometimes, the person you gave information to would abuse the account himself, thus rendering it useless to you. Other times, the person would post the information publicly and claim credit for getting the account. (p. 65)

Interestingly enough, such concerns prompted the formation of the "Inner Circle", an elite group of high level hackers who felt they could trust one another, and who shared information only with each other!