THE NEW COGNITIVE NEUROSCIENCES

Second Edition

Michael S. Gazzaniga, *Editor-in-Chief*

0-262-07195-9

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

87 The Cognitive Neuroscience of Social Reasoning

LEDA COSMIDES AND JOHN TOOBY

ABSTRACT Cognitive scientists need theoretical guidance that is grounded in something beyond intuition. They need evolutionary biology's "adaptationist program": a research strategy in which theories of adaptive function are key inferential tools, used to identify and investigate the design of evolved systems. Using research on how humans reason about social exchange, the authors will (1) illustrate how theories of adaptive function can generate detailed and highly testable hypotheses about the design of computational machines in the human mind and (2) review research that tests for the presence of these machines. This research suggests that the human computational architecture contains an expert system designed for reasoning about cooperation for mutual benefit, with a subroutine specialized for cheater detection.

Natural competences

Scientists have been dissecting the neural architecture of the human mind for several centuries. Dissecting its computational architecture has proven more difficult, however. Our natural competences–our abilities to see, to speak, to find someone beautiful, to reciprocate a favor, to fear disease, to fall in love, to initiate an attack, to experience moral outrage, to navigate a landscape, and myriad others–are possible only because there is a vast and heterogeneous array of complex computational machinery supporting and regulating these activities. But this machinery works so well that we do not even realize that it exists. Our intuitions blur our scientific vision. As a result, we have neglected to study some of the most interesting machinery in the human mind.

Theories of adaptive function are powerful lenses that allow one to see beyond one's intuitions. Aside from those properties acquired by chance or imposed by engineering constraint, the mind consists of a set of information-processing circuits that were designed by natural selection to solve adaptive problems that our ancestors faced, generation after generation. If we know what these problems were, we can seek mechanisms that are well engineered for solving them.

The exploration and definition of adaptive problems is a major activity in evolutionary biology. By combining results derived from mathematical modeling, comparative studies, behavioral ecology, paleoanthropology, and other fields, evolutionary biologists try to identify (1) what problems the mind was designed to solve, (2) why it was designed to solve *those* problems rather than other ones, and (3) what information was available in ancestral environments that a problem-solving mechanism could have used. These are the components of what Marr (1982) called a "computational theory" of an information-processing problem: a task analysis defining what a computational device does and why it does it.

Because there are multiple ways of achieving any solution, experiments are always needed to determine which algorithms and representations actually evolved to solve a particular problem. But the more precisely you can define the goal of processing–the more tightly you can constrain what would count as a solution–the more clearly you can see what a program capable of producing that solution would have to look like. The more constraints you can discover, the more the field of possible solutions is narrowed, and the more you can concentrate your experimental efforts on discriminating between viable hypotheses.

In this way, theories of adaptive problems can guide the search for the cognitive programs that solve them. Knowing what cognitive programs exist can, in turn, guide the search for their neural basis. To illustrate this approach, we will show how it guided a research program of our own on how people reason about social interactions.

Some of the most important adaptive problems our ancestors had to solve involved navigating the social world, and some of the best work in evolutionary biology is devoted to analyzing constraints on the evolution of mechanisms that solve these problems. Constructing computational theories from these constraints led us to suspect that the human cognitive architecture contains expert systems specialized for reasoning about the social

LEDA COSMIDES Center for Evolutionary Psychology and Department of Psychology, University of California, Santa Barbara, Calif.
JOHN TOOBY Center for Evolutionary Psychology and Department of Anthropology, University of California, Santa Barbara, Calif.

world. If these exist, then their inference procedures, representational primitives, and default assumptions should reflect the structure of adaptive problems that arose when our hominid ancestors interacted with one another. Our first task analysis was of the adaptive information-processing problems entailed by the human ability to engage in social exchange.

Social exchange and conditional reasoning

In categorizing social interactions, there are two basic consequences that humans can have on each other: helping or hurting, bestowing benefits or inflicting costs. Some social behavior is unconditional: One nurses an infant without asking it for a favor in return, for example. But most social acts are delivered conditionally. This creates a selection pressure for cognitive designs that can detect and understand social conditionals reliably, precisely, and economically (Cosmides, 1985, 1989; Cosmides and Tooby, 1989, 1992). Two major categories of social conditionals are social exchange and threat–conditional helping and conditional hurting–carried out by individuals or groups on individuals or groups. We initially focused on social exchange (for review, see Cosmides and Tooby, 1992). A social exchange involves a conditional of the approximate form: *If person A provides the requested benefit to or meets the requirement of person or group B, then B will provide the rationed benefit to A.* (Herein, a rule expressing this kind of agreement to cooperate will be referred to as a *social contract.*)

We elected to study reasoning about social exchange for several reasons:

1. Many aspects of the evolutionary theory of social exchange (sometimes called *cooperation, reciprocal altruism,* or *reciprocation*) are relatively well developed and unambiguous. Consequently, certain features of the functional logic of social exchange could be confidently relied on in constructing *a priori* hypotheses about the structure of the information-processing procedures that this activity requires.

2. Complex adaptations are constructed in response to evolutionarily long-enduring problems. Situations involving social exchange have constituted a long-enduring selection pressure on the hominid line: Evidence from primatology and paleoanthropology suggests that our ancestors have engaged in social exchange for at least several million years.

3. Social exchange appears to be an ancient, pervasive, and central part of human social life. The universality of a behavioral phenotype is not a *sufficient* condition for claiming that it was produced by a cognitive adaptation, but it is suggestive. As a behavioral phenotype, social exchange is as ubiquitous as the human heartbeat. The heartbeat is universal because the organ that generates it is everywhere the same. This is a parsimonious explanation for the universality of social exchange as well: the cognitive phenotype of the organ that generates it is everywhere the same. Like the heart, its development does not seem to require environmental conditions (social or otherwise) that are idiosyncratic or culturally contingent.

4. Social exchange is relatively rare across species, however. Many species have the ability to recognize patterns (as connectionist systems do) or change their behavior in response to rewards and punishments (see chapter 81 of this volume). Yet these abilities alone are insufficient for social exchange to emerge, despite the rewards it can produce. This suggests that social exchange behavior is generated by cognitive machinery specialized for that task.[1]

5. Finding procedures specialized for reasoning about social exchange would challenge a central assumption of the behavioral sciences: that the evolved architecture of the mind consists solely or predominantly of a small number of content-free, general-purpose mechanisms (Tooby and Cosmides, 1992).

Reasoning is among the most poorly understood areas in the cognitive sciences. Its study has been dominated by a pre-Darwinian view, championed by the British Empiricists and imported into the modern behavioral sciences in the form of the Standard Social Science Model (SSSM). According to this view, reasoning is accomplished by circuits designed to operate uniformly over every class of content (see chapter 81), and the mind has no content that was not derived from the perceptual data these circuits take as input. These circuits were thought to be few in number, content free, and general purpose, part of a hypothetical faculty that generates solutions to all problems: "general intelligence." Experiments were designed to reveal what computational procedures these circuits embodied; prime candidates were all-purpose heuristics and "rational" algorithms–ones that implement formal methods for inductive and deductive reasoning, such as Bayes's rule or the propositional calculus. These algorithms are jacks of all trades: Because they are content free, they can operate on information from any domain (their strength). They are also masters of none: To be content independent means that they lack any domain-specialized information that would lead to correct inferences in one domain but would not apply to others (their weakness).

This view of reasoning as a unitary faculty composed of content-free procedures is intuitively compelling to many people. But the discipline of asking what adaptive

information-processing problems our minds evolved to solve changes one's scientific intuitions/sensibilities. One begins to appreciate (1) the complexity of most adaptive information-processing problems; (2) that the evolved solution to these problems is usually machinery that is well engineered for the task; (3) that this machinery is usually specialized to fit the particular nature of the problem; and (4) that its evolved design must embody "knowledge" about problem-relevant aspects of the world.

The human computational architecture can be thought of as a collection of evolved problem-solvers. Some of these may indeed embody content-free formalisms from mathematics or logic, which can act on any domain and acquire all their specific content from perceptual data alone (Gigerenzer, 1991; Brase, Cosmides, and Tooby, 1998). But many evolved problem-solvers are expert systems, equipped with "crib sheets": inference procedures and assumptions that embody knowledge specific to a given problem domain. These generate correct (or, at least, adaptive) inferences that would not be warranted on the basis of perceptual data alone. For example, there currently is at least some evidence for the existence of inference systems that are specialized for reasoning about objects (Baillergeon, 1986; Spelke, 1990), physical causality (Brown, 1990; Leslie, 1994), number (Gallistel and Gelman, 1992; Wynn, 1992, 1995), the biological world (Atran, 1990; Hatano and Inagaki, 1994; Keil, 1994; Springer, 1992), the beliefs and motivations of other individuals (Baron-Cohen, 1995; Leslie, 1987; see also chapters 85 and 86), and social interactions (Cosmides and Tooby, 1992; Fiske, 1991). These domain-specific inference systems have a distinct advantage over domain-independent ones, akin to the difference between experts and novices: Experts can solve problems faster and more efficiently than novices because they already know a lot about the problem domain.

So what design features might one expect an expert system that is well engineered for reasoning about social exchange to have?

Design features predicted by the computational theory

The evolutionary analysis of social exchange parallels the economist's concept of trade. Sometimes known as "reciprocal altruism," social exchange is an "I'll scratch your back if you scratch mine" principle. Economists and evolutionary biologists had already explored constraints on the emergence or evolution of social exchange using game theory, modeling it as a repeated Prisoners' Dilemma. Based on these analyses, and on data from paleoanthropology and primatology, we developed a computational theory (*sensu* Marr, 1982) specifying design features that algorithms capable of satisfying these constraints would have to have. For example:

1. To discriminate social contracts from threats and other kinds of conditionals, the algorithms involved would have to be sensitive to the presence of benefits and costs and be able to recognize a well-formed social contract (for a grammar of social exchange, see Cosmides and Tooby, 1989). Social exchange is cooperation for *mutual benefit.* The presence of a benefit is crucial for a situation to be recognized as involving social exchange. The presence of a cost is not a necessary condition: providing a benefit may cause one to incur a cost, but it need not. There must, however, be algorithms that can assess relative benefits and costs, to provide input to decision rules that cause one to accept a social contract only when the benefits outweigh the costs.

2. The game theoretic analyses indicated that social exchange cannot evolve in a species or be sustained stably in a social group unless the cognitive machinery of the participants allows a potential cooperator to detect individuals who cheat, so that they can be excluded from future interactions in which they would exploit cooperators (Axelrod, 1984; Axelrod and Hamilton, 1981; Boyd, 1988; Trivers, 1971; Williams, 1966). In this context, a *cheater* is an individual who accepts a benefit without satisfying the requirements that provision of that benefit was made contingent upon. This definition does not map onto content-free definitions of violation found in the propositional calculus and in most other reasoning theories (e.g., Rips, 1994; Johnson-Laird and Byrne, 1991). A system capable of detecting cheaters would need to define the concept using contentful representational primitives, referring to illicitly taken *benefits.* The definition also is perspective dependent because the item or action that one party views as a benefit, the other views as a requirement. Given "If you give me your watch, I'll give you $10," you would have cheated me if you took my $10 but did not give me your watch; I would have cheated you if I had taken your watch without giving you the $10. This means that the system needs to be able to compute a cost-benefit representation from the perspective of each participant, and define cheating with respect to that perspective-relative representation.

In short, what counts as cheating is so content dependent that a detection mechanism equipped with a domain-general definition of violation would not be able to solve the problem of cheater detection. Hence, an expert system designed for conditional reasoning about social exchange should have a subroutine *specialized* for detecting cheaters.

TABLE 87.1

Computational machinery that governs reasoning about social contracts

Design features predicted (and established)

1. It includes inference procedures specialized for detecting cheaters.
2. The cheater detection procedures cannot detect violations that do not correspond to cheating (e.g., mistakes when no one profits from the violation).
3. The machinery operates even in situations that are unfamiliar and culturally alien.
4. The definition of cheating varies lawfully as a function of one's perspective.
5. The machinery is just as good at computing the cost-benefit representation of a social contract from the perspective of one party as from the perspective of another.
6. It cannot detect cheaters unless the rule has been assigned the cost-benefit representation of a social contract.
7. It translates the surface content of situations involving the contingent provision of benefits into representational primitives, such as "benefit," "cost," "obligation," "entitlement," "intentional," and "agent."
8. It imports these conceptual primitives, even when they are absent from the surface content.
9. It derives the implications specified by the computational theory, even when these are not valid inferences of the propositional calculus (e.g., "If you take the benefit, then you are obligated to pay the cost" implies "If you paid the cost, then you are entitled to take the benefit").
10. It does not include procedures specialized for detecting altruists (individuals who have paid costs but refused to accept the benefits to which they are therefore entitled).
11. It cannot solve problems drawn from other domains (e.g., it will not allow one to detect bluffs and double-crosses in situations of threat).
12. It appears to be neurologically isolable from more general reasoning abilities (e.g., it is unimpaired in schizophrenic patients who show other reasoning deficits; Maljkovic, 1987).
13. It appears to operate across a wide variety of cultures (including an indigenous population of hunter-horticulturists in the Ecuadorian Amazon; Sugiyama, Tooby, and Cosmides, 1995).

Alternative (by-product) hypotheses eliminated

1. That familiarity can explain the social contract effect.
2. That social contract content merely activates the rules of inference of the propositional calculus.
3. That social contract content merely promotes (for whatever reason) "clear thinking."
4. That permission schema theory can explain the social contract effect.
5. That any problem involving payoffs will elicit the detection of violations.
6. That a content-independent deontic logic can explain the effect.

Based on evidence reviewed in Cosmides and Tooby, 1992.

3. Algorithms regulating social exchange should be able to operate even over unfamiliar contents and situations. Unlike other primates, who exchange only a limited array of favors (e.g., grooming, food, protection), humans trade an almost unlimited variety of goods and services. Moreover, one needs to be able to interpret each new situation that arises–not merely ones that have occurred in the past. Thus, the algorithms should be able to operate properly even in unfamiliar situations, as long as they can be interpreted as involving the conditional provision of benefits. This means the representational format of the algorithms cannot be tied to specific items with a fixed exchange rate (e.g., one could imagine the reciprocation algorithms of vampire bats, who share regurgitated blood, to specify an exchange rate in blood volume). From the surface content of a situation, the algorithms should compute an abstract level of representation, with representational primitives such as $benefit_{agent\ 1}$, $requirement_{agent\ 1}$, *agent 1*, *agent 2*, $cost_{agent\ 2}$, and so forth.

4. In the context of social exchange, modals such as "must" and "may" should be interpreted deontically, as referring to obligation and entitlement (rather than to necessity and possibility). As a result, cheating is taking a benefit one is not entitled to. It does not matter where terms such as "benefit taken" or "requirement not met" fall in the logical structure of a rule. In addition, there are constraints on when one should punish cheating, and by how much (Cosmides, 1985; Cosmides and Tooby, 1989).

Such analyses provided a principled basis for generating detailed hypotheses about reasoning procedures that, because of their domain-specialized structure, would be well designed for detecting social conditionals involving exchange, interpreting their meaning, and successfully solving the inference problems they pose (table 87.1). These hypotheses were tested using standard methods from cognitive psychology, as described below.

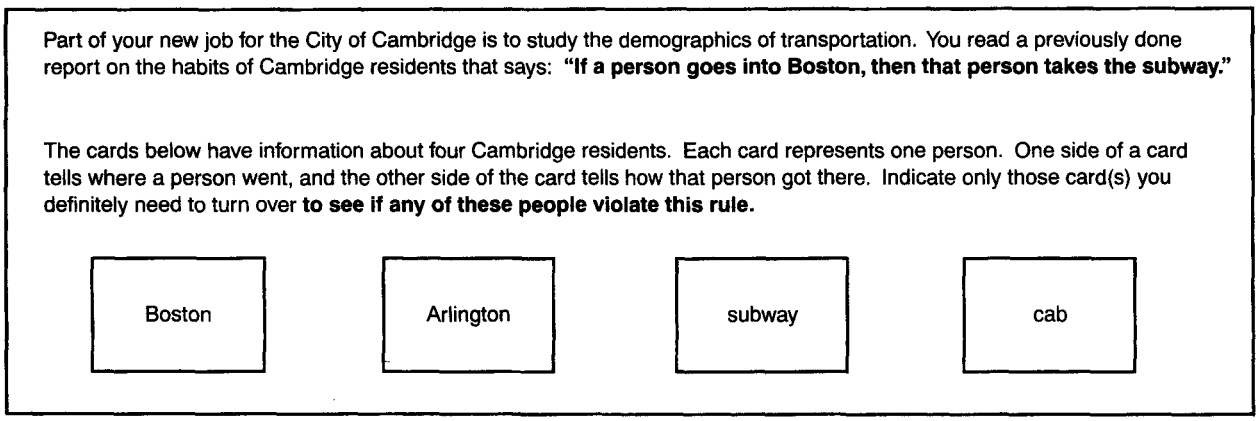

FIGURE 87.1 The Wason selection task (descriptive rule, familiar content). In a Wason selection task, there is always a rule of the form *If P then Q*, and four cards showing the values *P*, *not-P*, *Q*, and *not-Q* (respectively) on the side that the subject can see. From a logical point of view, only the combination of *P* and *not-Q* can violate this rule, so the correct answer is to check the *P* card (to see whether it has a *not-Q* on the back), the *not-Q* card (to see whether it has a *P* on the back), and no others. Few subjects answer correctly, however, when given problems with descriptive rules, such as the problem in figure 87.1.

Computational theories (or task analyses) are important because they specify a mechanism's adaptive function: the problem it was designed by natural selection to solve. They are central to any evolutionary investigation of the mind, for a simple reason. To show that an aspect of the phenotype is an adaptation, one needs to demonstrate a fit between form and function: One needs *design evidence*. There are now a number of experiments on human reasoning comparing performance on tasks in which a conditional rule either did or did not express a social contract. These experiments–some of which are described below–have provided evidence for a series of domain-specific effects predicted by our analysis of the adaptive problems that arise in social exchange. Social contracts activate content-*dependent* rules of inference that appear to be complexly specialized for processing information about this domain. These rules include subroutines that are specialized for solving a particular problem within that domain: cheater detection. The programs involved do not operate so as to detect potential altruists (individuals who pay costs but do not take benefits), nor are they activated in social contract situations in which errors would correspond to innocent mistakes rather than intentional cheating. Nor are they designed to solve problems drawn from domains other than social exchange; for example, they do not allow one to detect bluffs and double-crosses in situations of threat, nor do they allow one to detect when a safety rule has been violated. The pattern of results elicited by social exchange content is so distinctive that we believe reasoning in this domain is governed by computational units that are domain specific and functionally distinct: what we have called *social contract algorithms* (Cosmides, 1985, 1989; Cosmides and Tooby, 1992).

To help readers track which hypotheses are tested by each experiment, we will refer to the list in table 87.1. For example, D1 = design feature #1 (inference procedures specialized for detecting cheaters); B1 = byproduct hypothesis #1 (that familiarity can explain the social contract effect).

Tests with the Wason selection task

To test for the presence of the design features predicted, we used an experimental paradigm called the Wason selection task (Wason, 1966; Wason and Johnson-Laird, 1972). For more than 30 years, psychologists have been using this paradigm (which was orginally developed as a test of logical reasoning) to probe the structure of human reasoning mechanisms. In this task, the subject is asked to look for violations of a conditional rule of the form *If P then Q*. Consider the Wason selection task presented in figure 87.1.

From a logical point of view, the rule has been violated whenever someone goes to Boston without taking the subway. Hence, the logically correct answer is to turn over the *Boston* card (to see if this person took the subway) and the *cab* card (to see if the person taking the cab went to Boston). More generally, for a rule of the form *If P then Q*, one should turn over the cards that represent the values *P* (a true antecedent) and *not-Q* (a false consequent).

If the human mind develops reasoning procedures specialized for detecting logical violations of conditional rules, this would be intuitively obvious. But it is not. In general, fewer than 25% of subjects spontaneously make this response. Moreover, even formal training in logical reasoning does little to boost performance on descriptive

rules of this kind (Cheng et al., 1986; Wason and Johnson-Laird, 1972). Indeed, a large literature exists that shows that people are not very good at detecting logical violations of if-then rules in Wason selection tasks, *even when these rules deal with familiar content drawn from everyday life* (Manktelow and Evans, 1979; Wason, 1983).

The Wason selection task provided an ideal tool for testing hypotheses about reasoning specializations designed to operate on social conditionals, such as social exchanges, threats, permissions, obligations, and so on, because (1) it tests reasoning about conditional rules, (2) the task structure remains constant while the content of the rule is changed, (3) content effects are easily elicited, and (4) there was already a body of existing experimental results against which performance on new content domains could be compared.

For example, to show that people who ordinarily cannot detect violations of conditional rules can do so when that violation represents cheating on a social contract would constitute initial support for the view that people have cognitive adaptations specialized for detecting cheaters in situations of social exchange. To find that violations of conditional rules are spontaneously detected when they represent bluffing on a threat would, for similar reasons, support the view that people have reasoning procedures specialized for analyzing threats. Our general research plan has been to use subjects' inability to spontaneously detect violations of conditionals expressing a wide variety of contents as a comparative baseline against which to detect the presence of performance-boosting reasoning specializations. By seeing which content-manipulations switch on or off high performance, the boundaries of the domains within which reasoning specializations successfully operate can be mapped.

The results of these investigations were striking. People who ordinarily cannot detect violations of if-then rules can do so easily and accurately when that violation represents cheating in a situation of social exchange (Cosmides, 1985, 1989; Cosmides and Tooby, 1989, 1992). Given a rule of the general form, "If you take benefit B, then you must satisfy requirement R," subjects choose the *benefit accepted* card and the *requirement not met* card–the cards that represent potential cheaters. The adaptively correct answer is immediately obvious to almost all subjects, who commonly experience a "pop out" effect. No formal training is needed. Whenever the content of a problem asks one to look for cheaters in a social exchange, subjects experience the problem as simple to solve, and their performance jumps dramatically. In general, 65% to 80% of subjects get it right, the highest performance found for a task of this kind (supports D1).

This is true for familiar social contracts, such as, "If a person drinks beer, then that person must be over 19

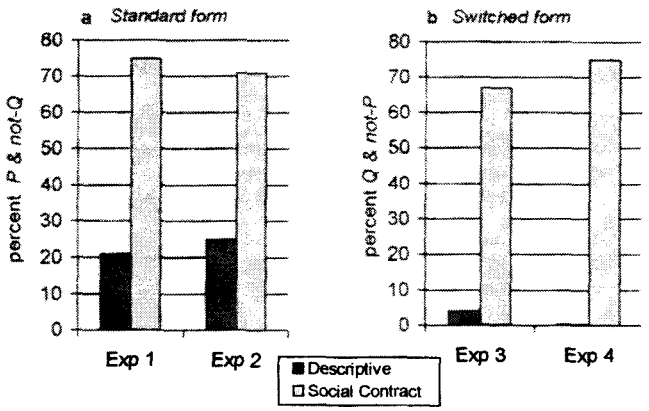

FIGURE 87.2 Detecting violations of unfamiliar conditional rules: social contracts versus descriptive rules. In these experiments, the same, unfamiliar rule was embedded either in a story that caused it to be interpreted as a social contract or in a story that caused it to be interpreted as a rule describing some state of the world. For social contracts, the correct answer is always to pick the *benefit accepted* card and the *requirement not met* card. (A) For standard social contracts, these correspond to the logical categories *P* and *not-Q*. *P and not-Q* also happens to be the logically correct answer. More than 70% of subjects chose these cards for the social contracts, but fewer than 25% chose them for the matching descriptive rules. (B) For switched social contracts, the *benefit accepted* and *requirement not met* cards correspond to the logical categories *Q* and *not-P*. This is not a logically correct response. Nevertheless, approximately 70% of subjects chose it for the social contracts; virtually no one chose it for the matching descriptive rule.

years old" (Griggs and Cox, 1982; Cosmides, 1985). According to our computational theory, however, performance also should be high for unfamiliar ones–such as, "If a man eats cassava root, then he must have a tattoo on his face," where cassava root is portrayed as a highly desirable aphrodisiac and having a facial tattoo is a sign of being married. As figure 87.2A shows, this is true. Subjects choose the *benefit accepted* card (e.g., "ate cassava root") and the *requirement not met* card (e.g., "no tattoo") for any social conditional that can be interpreted as a social contract and in which looking for violations can be interpreted as looking for cheaters (supports D1, D3, D7, D8; disconfirms B1; Cosmides, 1985, 1989; Gigerenzer and Hug, 1992; Platt and Griggs, 1993). Indeed, familiarity did not help at all. Cosmides (1985) found that performance was just as high on unfamiliar as on familiar social contracts–an uncomfortable result for any explanation that invokes domain-general learning, which depends on familiarity and repetition.

From a domain-general, formal view, investigating men eating cassava root and men without tattoos is logi-

cally equivalent to investigating people going to Boston and people taking cabs. But everywhere it has been tested (adults in the United States, United Kingdom, Germany, Italy, France, Hong-Kong, Japan; schoolchildren in Ecuador; Shiwiar hunter-horticulturists in the the Ecuadorian Amazon (Sugiyama, Tooby, and Cosmides, 1995), people do not treat social exchange problems as equivalent to other kinds of reasoning problems. Their minds distinguish social exchange contents and reason as if they were translating these situations into representational primitives such as "benefit," "cost," "obligation," "entitlement," "intentional," and "agent" (Cheng and Holyoak, 1985; Cosmides, 1989; Platt and Griggs, 1993; supports D13). Indeed, the relevant inference procedures are not activated unless the subject has represented the situation as one in which one is entitled to a benefit only if one has satisfied a requirement.

Do social contracts simply activate content-free logical rules?

The procedures activated by social contract rules do not behave as if they were designed to detect *logical* violations per se; instead, they prompt choices that track what would be useful for detecting cheaters, regardless of whether this happens to correspond to the logically correct selections. For example, by switching the order of requirement and benefit within the if-then structure of the rule (figure 87.3), one can elicit responses that are functionally correct from the point of view of cheater detection, but logically incorrect. Subjects choose the *benefit accepted* card and the *requirement not met* card–the adaptively correct response if one is looking for cheaters–*no matter what logical category these cards fall into* (figure 87.2B; supports D9, D8; disconfirms B2, B3, B1) This means that cheater detection is not accomplished by procedures embodying the content-free rules of logical inference. If it were, then subjects would choose the logically correct response–*P and not-Q*–even on switched rules, where these represent the *requirement met* card (P) and the *benefit not accepted* card (*not-Q*). Yet these represent people who cannot possibly have cheated. A person who has met a requirement without accepting the benefit this entitles one to is either an altruist or a fool, but not a cheater.

That content-free rules of logic are not responsible for cheater detection was demonstrated in a different way by Gigerenzer and Hug (1992) in experiments designed to test the prediction that what counts as cheating should depend on one's perspective (D4, D5, B2, B4). Subjects were asked to look for violations of rules such as, "If a previous employee gets a pension from a firm, then that person must have worked for the firm for at least 10 years." For half of them, the surrounding story cued the subjects into the role of the employer, whereas for the other half, it cued them into the role of an employee.

If social contracts activate rules for detecting logical violations, then this manipulation should make no difference. Perspectives, such as that of employer versus employee, play no role in formal logic: To detect a logical violation, the *P* card ("employee got the pension") and the *not-Q* card ("employee worked for less than 10 years") should be chosen, no matter whose perspective you have taken. But in social contract theory, perspective matters. Because these two cards represent instances in which the employer may have been cheated, a cheater detection subroutine should choose them in the employer condition. But in the employee condition, the same cheater detection procedures ought to draw attention to situations in which an *employee* might have been cheated, that is, the "employee worked for more than 10 years" card (Q) and "employee got no pension" card (*not-P*). In the employee condition, *Q and not-P* is logically incorrect, but adaptively correct.

The results confirmed the social contract prediction. In the employer condition, 73% of subjects chose *P and not-Q* (which is both logically and adaptively correct), and less than 1% chose *Q and not-P* (which is both logically and adaptively incorrect). But in the employee condition, 66% chose *Q and not-P*–which is logically incorrect but adaptively correct–and only 8% chose *P and not-Q*–which is logically correct, but adaptively incorrect (supports D4, D7, D9; disconfirms B2, B4). Furthermore, the percent of subjects choosing the correct cheater detection response did not differ significantly in these two conditions (73% vs. 66%), indicating that it was just as easy for subjects to compute a cost-benefit representation from the perspective of the employer as from that of the employee (supports D5).

Are cheater detection procedures part of "general intelligence"?

Data from Maljkovic (1987) show that the ability to detect cheaters can remain intact even in individuals suffering large impairments in their more general intellectual functioning. Maljkovic tested the reasoning of patients suffering from positive symptoms of schizophrenia, comparing their performance to that of hospitalized controls. The schizophrenic patients were indeed impaired on more general tests of logical reasoning, in a way typical of individuals with frontal lobe dysfunction. But their ability to detect cheaters on Wason selection tasks was unimpaired (supports D1, D12; disconfirms

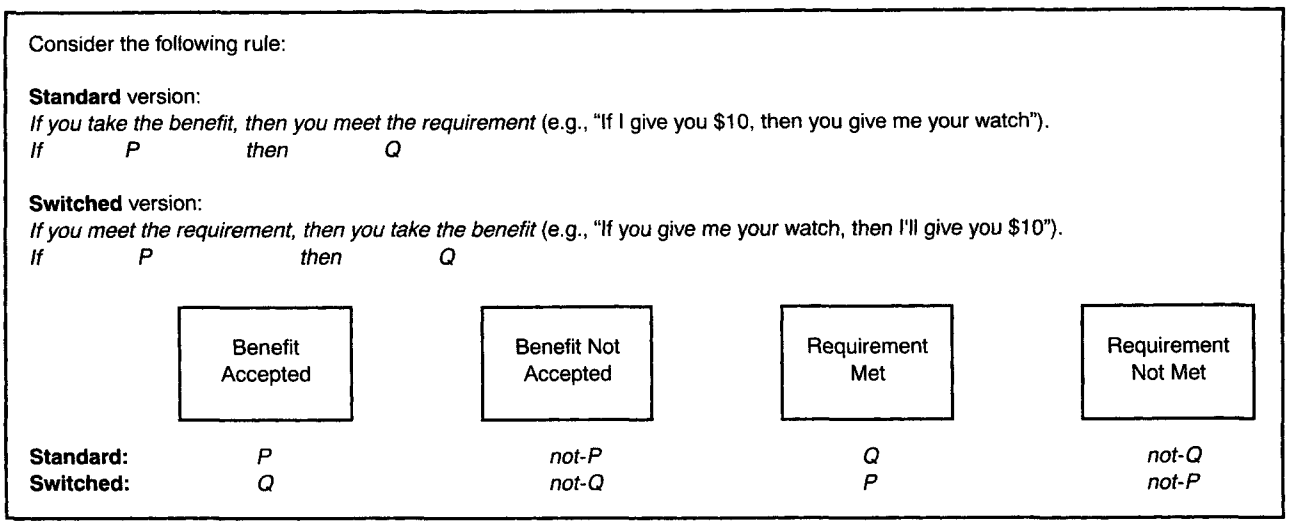

FIGURE 87.3 Generic structure of a social contract. A cheater detection subroutine always would cause one to look at instances in which the potential violator has taken the benefit and not met the requirement. But whether this corresponds to the logically correct answer (*P and not-Q*) depends on exactly how the exchange was expressed. The same social contract has been offered to you, whether the offerer says, "If I give you $10, then you give me your watch" (standard) or "If you give me your watch, I'll give you $10" (switched). But the benefit to you (getting the $10) is in the *P* clause in the standard version and in the *Q* clause in the switched version. Likewise, your not meeting the requirement (not giving me the watch) would correspond to the logical category *not-Q* in the standard version and *not-P* in the switched version. By always choosing the *benefit accepted* and *requirement not met* cards, a cheater detection procedure would cause one to choose *P and not-Q* in the standard version–a logically correct response–and *Q and not-P* in the switched version–a logically incorrect response. By testing switched social contracts, one can see that the reasoning procedures activated cause one to detect cheaters, not logical violations.

B1, B3). This is all the more remarkable given that schizophrenics usually show impairments on virtually any test of intellectual functioning that they are given (McKenna, Clare, and Baddeley, 1995). Maljkovic argues that a dissociation of this kind is what one would expect if social exchange, which is a long-enduring adaptive problem, is generated by mechanisms in more evolutionarily ancient parts of the brain than the frontal lobes. Whether this conjecture is true or not, her results indicate that the algorithms responsible for cheater detection are different from those responsible for performance on the more general logical reasoning tasks these patients were given.

Are these effects produced by permission schemas?

The human cognitive phenotype has many features that appear to be complexly specialized for solving the adaptive problems that arise in social exchange. But demonstrating this is not sufficient for claiming that these features are cognitive adaptations *for* social exchange. One also needs to show that these features are not more parsimoniously explained as the by-product of mechanisms designed to solve some other adaptive problem or class of problems.

TABLE 87.2
The permission schema is composed of four production rules

1. Rule 1: If the action is to be taken, then the precondition must be satisfied.
2. Rule 2: If the action is not to be taken, then the precondition need not be satisfied.
3. Rule 3: If the precondition is satisfied, then the action may be taken.
4. Rule 4: If the precondition is not satisfied, then the action must not be taken.

From Cheng and Holyoak, 1985.

For example, Cheng and Holyoak (1985, 1989) also invoke content-dependent computational mechanisms to explain reasoning performance that varies across domains. But they attribute performance on social contract rules to the operation of a permission schema (and/or an obligation schema; these do not lead to different predictions on the kinds of rules usually tested; see Cosmides, 1989), which operates over a larger class of problems. They propose that this schema consists of four production rules (table 87.2) and that their scope is any permission rule, that is, any conditional rule to which the subject assigns the following abstract representation: "If action A is to be taken, then precondition P must be satisfied." All social contracts are permission rules, but not all permission rules are social contracts. The conceptual

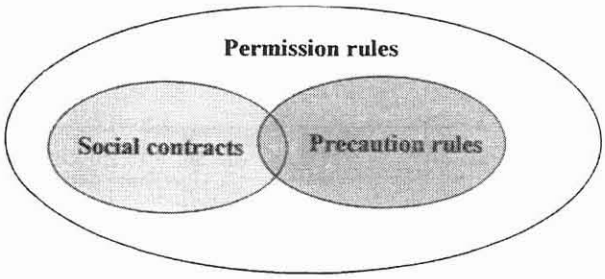

FIGURE 87.4 All social contracts are permission rules; all precaution rules are permission rules; but not all permission rules are social contracts or precautions. Many permission rules that have been tested fall into the white area (neither social contracts nor precautions): these do not elicit the high levels of performance that social contracts and precaution rules do. This argues against permission schema theory. By testing patients with focal brain damage and through priming studies, one can see whether reasoning about social contracts and precautions can be dissociated. These methods allow one to determine whether reasoning about social contracts and precaution rules is generated by one permission schema, or by two separate domain-specific mechanisms.

primitives of a permission schema have a larger scope than those of social contract algorithms. For example, a "benefit taken" is a kind of "action taken," and a "cost paid" (i.e., a benefit offered in exchange) is a kind of "precondition satisfied." They take evidence that people are good at detecting violations of precaution rules–rules of the form, "If hazardous action H is taken, then precaution P must be met"–as evidence for their hypothesis (on precautions, see Manktelow and Over, 1988, 1990). After all, a precaution rule is a kind of permission rule, but it is not a kind of social contract. We, however, have hypothesized that reasoning about precaution rules is governed by a functionally specialized inference system that differs from social contract algorithms and operates independently of them (Cosmides and Tooby, 1992, 1997; Fiddick, 1998; Fiddick, Cosmides, and Tooby, 1995; figure 87.4).

In other words, there are two competing proposals for how the computational architecture that causes reasoning in these domains should be dissected. Several lines of evidence speak to these competing claims.

ARE BENEFITS NECESSARY? According to the grammar of social exchange, a rule is not a social contract unless it contains a *benefit to be taken.* Transformations of input should not matter, as long as the subject continues to represent an action or state of affairs as beneficial to the potential violator and the violator as illicitly obtaining this benefit. The corresponding argument of the permission schema–an *action to be taken*–has a larger scope: Not all "actions taken" are "benefits taken." If this construal of the rule's representational structure is correct, then the behavior of the reasoning system should be invariant over transformations of input that preserve it. But it is not. For example, consider two rules: (1) "If one goes out at night, then one must tie a small piece of red volcanic rock around one's ankle" and (2) "If one takes out the garbage at night, then one must tie a small piece of red volcanic rock around one's ankle." Most undergraduate subjects perceive the action to be taken in (1)–going out at night–as a benefit, and 80% of them answered correctly. But when one substitutes a different action–taking out the garbage–into the same place in the argument structure, then performance drops to 44% (supports D6, D7; disconfirms B4, B6; Cosmides and Tooby, 1992). This transformation of input preserves the *action to be taken* representational structure, but it does not preserve the *benefit to be taken* representational structure–most people think of taking out the garbage as a chore, not a benefit. If the syntax of the permission schema were correct, then performance should be invariant over this transformation. But a drop in performance is expected if the syntax of the social contract algorithms is correct.

We have been doing similar experiments with precaution rules (e.g., "If you make poison darts, then you must wear rubber gloves."). All precaution rules are permission rules (but not all permission rules are precaution rules). We have been finding that the degree of hazard does not affect performance, but the nature of the precaution does–even though all the *precautions taken* are instances of *preconditions satisfied.* Performance drops when the precaution is not perceived as a good safeguard given the hazard specified (Rutherford, Tooby, and Cosmides, 1996). This is what one would expect if the syntax of the rules governing reasoning in this domain take representations such as *facing a hazard* and *precaution taken*; it is not what one would expect if the representations were *action taken* and *precondition satisfied.*

DOES THE VIOLATION HAVE TO BE CHEATING? By hypothesis, social contract algorithms contain certain conceptual primitives that the permission schema lacks. For example, *cheating* is taking a benefit that one is not entitled to; we have proposed that social contract algorithms have procedures that are specialized for detecting *cheaters.* This conceptual primitive plays no role in the operation of the permission schema. For this schema, whenever the action has been taken but the precondition has not been satisfied, a *violation* has occurred. People should be good at detecting violations, whether that violation counts as cheating (the benefit has been illicitly taken by the violator) or a mistake (the violator does not get the benefit stipulated in the rule).

Given the same social contract rule, one can manipulate contextual factors to change the nature of the violation from cheating to a mistake. When we did this, performance changed radically, from 68% correct in the cheating condition to 27% correct in the mistake condition (supports D2; disconfirms B1-B6). Gigerenzer and Hug (1992) found the same drop in response to a similar context manipulation.

In bargaining games, experimental economists have found that subjects are twice as likely to punish defections when it is clear that the defector intended to cheat as when the defector is a novice who might have simply made a mistake (Hoffman, McCabe, and Smith, 1997). This provides interesting convergent evidence, using entirely different methods, for a conceptual distinction between mistakes and cheating, where intentionality also plays a role.

IS REASONING ABOUT SOCIAL CONTRACTS AND PRECAUTIONS GENERATED BY ONE MECHANISM OR TWO? If reasoning about social contracts and precautions is caused by one and the same mechanism–a permission schema–then neurological damage to this schema should lower performance on both rules equally. But if reasoning about these two domains is caused by two, functionally distinct mechanisms, then one could imagine neurological damage to the social contract algorithms that leaves the precaution mechanisms unimpaired, and vice versa. Stone and colleagues (Stone, Cosmides, and Tooby, 1996; Stone, Cosmides, and Tooby, forthcoming; Stone et al., 1997) tested R. M., a patient with bilateral damage to his orbitofrontal and anterior temporal cortex, as well as to the left amygdala, on social contract and precaution problems that had been matched for difficulty on normal controls (who got 71% and 73% correct, respectively). R. M.'s performance on the precaution problems was 70% correct: equivalent to that of the normal controls. In contrast, his performance on the social contract problems was only 39% correct. This is a marked impairment, whether compared to the normal controls or to his own performance on the precaution problems (figure 87.5). R. M.'s difference score–his percent correct for precautions minus his percent correct for social contracts–was 31 percentage points. In contrast, the difference scores for individual normal subjects were all close to zero (mean = 0.14 percentage points). If reasoning on both social contracts and precautions were caused by a single mechanism–whether a permission schema or anything else–then one would not be able to find individuals who perform well on one class of content but not on the other. This pattern of results is best explained by the hypothesis that reasoning about these two types of content is governed by two separate mechanisms.

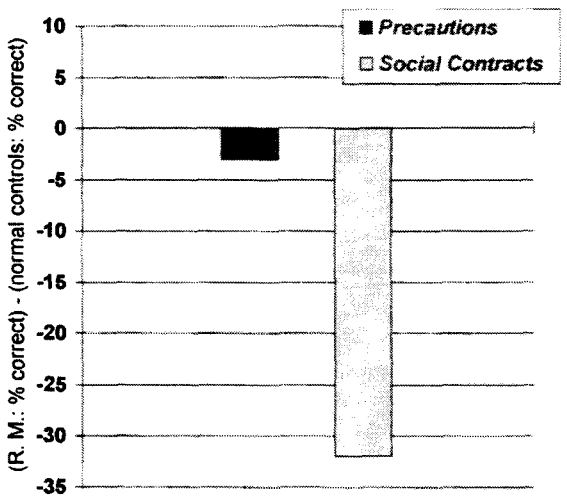

FIGURE 87.5 Performance of R.M. versus normal controls: precautions versus social contracts. Selective impairment of social contract reasoning in R. M., a patient with bilateral damage to the orbitofrontal and anterior temporal lobes and damage to the left amygdala. R. M. reasons normally when asked to look for violations of precaution rules but has difficulty when asked to look for (logically isomorphic) violations of social contracts (i.e., detecting cheaters). The *y*-axis plots the difference between R. M.'s percent correct and the mean performance of 37 normal controls. (Data from Stone, Cosmides, and Tooby, 1996.)

Although tests of this kind cannot conclusively establish the anatomical location of a mechanism, tests with other patients suggest that amygdalar damage was important in creating this selective deficit. Stone and associates tested two other patients who had no damage to the amygdala. One had bilateral orbitofrontal and anterior temporal damage; the other had bilateral anterior temporal damage, but no orbitofrontal damage. Neither patient exhibited a selective deficit; indeed, both scored extremely high on both classes of problems.

Convergent evidence for the single dissociation found by Stone and colleagues comes from a study by Fiddick and associates (Fiddick, Cosmides, and Tooby, 1995; Fiddick, 1998). Using a priming paradigm, they produced a functional dissociation in reasoning in normal, brain-intact subjects. Indeed, they were able to produce a *double* dissociation between social contract and precaution reasoning. More specifically, they found that: (1) when the problem used as a prime involved a clear social contract rule, performance on the target problem–an ambiguous social contract–increased. Moreover, this was due to the activation of social contract categories, not logical ones: when the prime was a switched social contract, in which the correct cheater

detection answer is *not* the logically correct answer (see figure 87.3), subjects matched their answers on the target to the prime's benefit/requirement categories, not its logical categories. (2) When the prime was a clear precaution rule, performance on an ambiguous precaution target increased. Most importantly, (3) these effects were caused by the operation of two mechanisms, rather than one: the precaution prime produced little or no increase in performance on an ambiguous social contract target; similarly, the social contract prime produced little or no increase in performance on an ambiguous precaution target.

This should not happen if permission schema theory were correct. In that view, it should not matter which rule is used as a prime because the only way in which social contracts and precautions can affect the interpretation of ambiguous rules is through activating the more general permission schema. Because both types of rules strongly activate this schema, an ambiguous target should be primed equally by either one.

Conclusion

Many cognitive scientists believe that theories of adaptive function are an explanatory luxury–fanciful, unfalsifiable post-hoc speculations that one indulges in at the end of a project, after the hard work of experimentation has been done. Nothing could be farther from the truth. By using a computational theory specifying the adaptive problems entailed by social exchange, we and our colleagues were able to predict, in advance, that certain very subtle changes in content and context would produce dramatic differences in how people reason. Without this theory, it is unlikely that this very precise series of effects would have been found. Even if someone stumbled upon a few of them, it is unlikely that their significance would have been recognized. Indeed, the situation would be similar to that in 1982, when we started this work: cognitive scientists could not understand why some familiar content–such as the drinking-age problem–produced "logical" reasoning whereas other familiar content–such as the transportation problem–did not (Griggs and Cox, 1982). By applying the adaptationist program, we were able to explain what was already known and to discover design features that no one had thought to test for before.

To isolate a functionally organized mechanism within a complex system, one needs a theory of what function that mechanism was designed to perform. The goal of cognitive neuroscience is to dissect the computational architecture of the human mind into functional units. The adaptationist program is cognitive neuroscience's best hope for achieving this goal.

ACKNOWLEDGMENTS The authors gratefully acknowledge the financial support of the James S. McDonnell Foundation, the National Science Foundation (NSF Grant BNS9157-449 to John Tooby), and a Research Across Disciplines grant (Evolution and the Social Mind) from the UCSB office of Research.

NOTE

1. Its relative rarity also suggests that this machinery is unlikely to evolve unless certain other cognitive capacities–components of a theory of mind, perhaps–are already in place.

REFERENCES

ATRAN, S., 1990. *The Cognitive Foundations of Natural History.* New York: Cambridge University Press.

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

AXELROD, R., and W. D. HAMILTON, 1981. The evolution of cooperation. *Science* 221:1390–1396.

BAILLARGEON, R., 1986. Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month old infants. *Cognition* 23:21–41.

BARON-COHEN, S., 1995. *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge, Mass.: MIT Press.

BOYD, R., 1988. Is the repeated prisoner's dilemma a good model of reciprocal altruism? *Ethol. Sociobiol.* 9:211–222.

BRASE, G., L. COSMIDES, and J. TOOBY, 1998. Individuation, counting, and statistical inference: The role of frequency and whole object representations in judgment under uncertainty. *J. Psychol. Gen.* 127:1–19.

BROWN, A., 1990. Domain-specific principles affect learning and transfer in children. *Cogn. Sci.* 14:107–133.

CHENG, P., and K. HOLYOAK, 1985. Pragmatic reasoning schemas. *Cogn. Psychol.* 17:391–416.

CHENG, P., and K. HOLYOAK, 1989. On the natural selection of reasoning theories. *Cognition* 33:285–313.

CHENG, P., K. HOLYOAK, R. NISBETT, and L. OLIVER, 1986. Pragmatic versus syntactic approaches to training deductive reasoning. *Cogn. Psychol.* 18:293–328.

COSMIDES, L., 1985. *Deduction or Darwinian algorithms? An Explanation of the "Elusive" Content Effect on the Wason Selection Task.* Doctoral dissertation, Department of Psychology, Harvard University, Cambridge, Mass., University Microfilms, #86 02206.

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

COSMIDES, L., and J. TOOBY, 1989. Evolutionary psychology and the generation of culture, part II. Case study: A computational theory of social exchange. *Ethol. Sociobiol.* 10:51–97.

COSMIDES, L., and J. TOOBY, 1992. Cognitive adaptations for social exchange. In *The Adapted Mind,* J. Barkow, L. Cosmides, and J. Tooby, eds. New York: Oxford University Press, pp. 163–228.

COSMIDES, L., and J. TOOBY, 1997. Dissecting the computational architecture of social inference mechanisms. In *Characterizing Human Psychological Adaptations* (Ciba Foundation Symposium #208). Chichester: Wiley, pp. 132–156.

FIDDICK, L., 1988. *The Deal and the Danger: An Evolutionary Analysis of Deonitic Reasoning.* Doctoral dissertation, University of California, Santa Barbara, Calif.

FIDDICK, L., L. COSMIDES, and J. TOOBY, 1995. *Priming Darwinian Algorithms: Converging Lines of Evidence for Domain-Specific Inference Modules.* Presented at the Seventh Annual meeting of the Human Behavior and Evolution Society, Santa Barbara, Calif., June 28–July 2, 1995.

FISKE, A., 1991. *Structures of Social Life: The Four Elementary Forms of Human Relations.* New York: Free Press.

GALLISTEL, C., and R. GELMAN, 1992. Preverbal and verbal counting and computation. *Cognition* 44:43–74.

GIGERENZER, G., 1991. How to make cognitive illusions disappear: Beyond heuristics and biases. *Eur. Rev. Soc. Psychol.* 2:83–115.

GIGERENZER, G., and K. HUG, 1992. Domain-specific reasoning: Social contracts, cheating and perspective change. *Cognition* 43:127–171.

GRIGGS, R., and J. COX, 1982. The elusive thematic-materials effect in Wason's selection task. *Br. J. Psychol.* 73:407–420.

HATANO, G., and K. INAGAKI, 1994. Young children's naive theory of biology. *Cognition* 50:171–188.

HOFFMAN, E., MCCABE, K., and SMITH, V., 1997. Behavioral foundations of reciprocity: Experimental economics and evolutionary psychology. Economic Science Laboratory, University of Arizona. To appear in *Economic Inquiry.*

JOHNSON-LAIRD, P., and R. BYRNE, 1991. *Deduction.* Hillsdale, N.J.: Erlbaum.

KEIL, F., 1994. The birth and nurturance of concepts by domain: The origins of concepts of living things. In *Mapping the Mind: Domain Specificity and Culture,* L. Hirschfeld and S. Gelman, eds. New York: Cambridge University Press, pp. 234–254.

LESLIE, A., 1987. Pretense and representation: The origins of "theory of mind." *Psychol. Rev.* 94:412–426.

LESLIE, A., 1994. ToMM, ToBY, and agency: Core architecture and domain specificity. In *Mapping the Mind: Domain Specificity in Cognition and Culture,* L. Hirschfeld and S. Gelman, eds. New York: Cambridge University Press, pp. 119–148.

MALJKOVIC, V., 1987. *Reasoning in Evolutionarily Important Domains and Schizophrenia: Dissociation Between Content-Dependent and Content Independent Reasoning.* Unpublished undergraduate honors thesis, Department of Psychology, Harvard University, Cambridge, Mass.

MANKTELOW, K., and D. OVER, 1988. *Sentences, Stories, Scenarios, and the Selection Task.* Presented at the First International Conference on Thinking. Plymouth, U.K. July 1988.

MANKTELOW, K., and D. OVER, 1990. Deontic thought and the selection task. In *Lines of Thinking,* Vol. 1, K. J. Gilhooly, M. T. G. Keane, R. H. Logie, and G. Erdos, eds. London: Wiley.

MANKTELOW, K. I., and J. ST. B. T. EVANS, 1979. Facilitation of reasoning by realism: Effect or non-effect? *Br. J. Psychol.* 70:477–488.

MARR, D., 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: Freeman.

MCKENNA, P., L. CLARE, and A. BADDELEY, 1995. Schizophrenia. In *Handbook of Memory Disorders,* A. D. Baddeley, B. A. Wilson, and F. N. Watts, eds. New York: Wiley.

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

RIPS, L., 1994. *The Psychology of Proof.* Cambridge, Mass.: MIT Press.

RUTHERFORD, M., J. TOOBY, and L. COSMIDES, 1996. *Adaptive Sex Differences in Reasoning About Self-Defense.* Presented at the Eighth Annual Meeting of the Human Behavior and Evolution Society, Northwestern University, Ill., June 26–30, 1996.

SPELKE, E., 1990. Principles of object perception. *Cogn. Sci.* 14:29–56.

SPRINGER, K., 1992. Children's awareness of the implications of biological kinship. *Child Dev.* 63:950–959.

STONE, V. E., S. BARON-COHEN, L. COSMIDES, J. TOOBY, and R. T. KNIGHT, 1997. Selective impairment of social inference abilities following orbitofrontal cortex damage. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society,* M. G. Shafto and P. Langley, eds. London: Lawrence Erlbaum, p. 1062.

STONE, V., L. COSMIDES, and J. TOOBY, 1996. *Selective Impairment of Cheater Detection: Neurological Evidence for Adaptive Specialization.* Presented at the Eighth Annual Meeting of the Human Behavior and Evolution Society, Northwestern University, Ill., June 26–30, 1996.

STONE, V., L. COSMIDES, and J. TOOBY, forthcoming. Selective impairment of social reasoning in an orbitofrontal and anterior temporal patient.

SUGIYAMA, L., J. TOOBY, and L. COSMIDES, 1995. *Cross-Cultural Evidence of Cognitive Adaptations for Social Exchange among the Shiwiar of Ecuadorian Amazonia.* Presented at the Seventh Annual Meetings of the Human Behavior and Evolution Society, University of California, Santa Barbara. Calif., June 28–July 2, 1995.

TOOBY, J., and L. COSMIDES, 1992. The psychological foundations of culture. In *The Adapted Mind,* J. Barkow, L. Cosmides, and J. Tooby, eds. New York: Oxford University Press, pp. 19–36.

TRIVERS, R., 1971. The evolution of reciprocal altruism. *Q. Rev. Biol.* 46 33–57.

WASON, P., 1983. Realism and rationality in the selection task. In *Thinking and Reasoning: Psychological Approaches,* J. St. B. T. Evans, ed. London: Routledge, pp. 44–75.

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

WASON, P., and P. JOHNSON-LAIRD, 1972. *The Psychology of Reasoning: Structure and content.* Cambridge, Mass.: Harvard University Press.

WILLIAMS, G., 1966. *Adaptation and Natural Selection.* Princeton: Princeton University Press.

WYNN, K., 1992. Addition and subtraction by human infants. *Nature* 358:749–750.

WYNN, K., 1995. Origins of numerical knowledge. *Math. Cogn.* 1:35–60.