Cognition, 31 (1989) 187–276

The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task*

LEDA COSMIDES

Stanford University

Received October 1987, final revision accepted October 1988

Abstract

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.

In order to successfully engage in social exchange—cooperation between two or more individuals for mutual benefit—humans must be able to solve a number of complex computational problems, and do so with special efficiency. Following Marr (1982), Cosmides (1985) and Cosmides and Tooby (1989) used evolutionary principles to develop a computational theory of these adaptive problems. Specific hypotheses concerning the structure of the algorithms that govern how humans reason about social exchange were derived from this computational theory. This article presents a series of experiments designed to test these hypotheses, using the Wason selection task, a test of logical reasoning. Part I reports experiments testing social exchange theory against the availability theories of reasoning; Part II reports experiments testing it against Cheng and Holyoak's (1985) permission schema theory. The experimental design included eight critical tests designed to choose between social exchange theory and these other two families of theories; the results of all eight tests support social exchange theory. The hypothesis that the human mind includes cognitive processes specialized for reasoning about social exchange predicts the content effects

---

*The theoretical perspective informing this paper was developed equally by John Tooby and myself, and will appear jointly elsewhere. Also, I thank him deeply for his astute and insightful criticism throughout. I am very grateful to Peter Wason for his timely assistance at the beginning of this project; to David Buss, Martin Daly, Irven DeVore, Paul Harvey, Richard J. Herrnstein, Stephen M. Kosslyn, Kenneth I. Manktelow, David E. Over, Dan Sperber, Valerie Stone, Sheldon H. White, Margo Wilson, and two anonymous reviewers for their excellent advice and comments; to Nasia Cosmides, Elena Eisenman, Mary Gomes, Naomi Rustomjee, Larissa Shyjan and Todd Truesdale for their kind help in conducting experiments, and to Roger Shepard for the intellectual stimulation and generous support he has given me. The experiments of Part I were supported by Harvard University; those of Part II were supported by NSF grant #BNS 85-11685 to Roger Shepard. Requests for reprints should be addressed to: Leda Cosmides, Department of Psychology, Bldg. 420, Stanford University, Stanford, CA 94305, U.S.A.

Editor's Note: This paper has won the 1988 American Association for the Advancement of Science Prize for Behavioral Science Research.

found in these experiments, and parsimoniously explains those that have already been reported in the literature. The implications of this line of research for a modular view of human reasoning are discussed, as well as the utility of evolutionary biology in the development of computational theories.

Introduction

Computational theories of adaptive problems

Even if they have not paid much attention to the fact, cognitive psychologists have always known that the human mind is not merely a computational system with the design features of a modern computer, but a biological system "designed" by the organizing forces of evolution. This means that the innate information-processing mechanisms that comprise the human mind were not designed to solve arbitrary tasks, but are, instead, *adaptations:* mechanisms designed to solve the specific biological problems posed by the physical, ecological and social environments encountered by our ancestors during the course of human evolution. However, most cognitive psychologists are not fully aware of just how useful these simple facts can be in the experimental investigation of human information-processing mechanisms.

The analytical tool through which evolutionary biology becomes useful to cognitive psychologists is the *computational theory*, as Marr (1982), for example, has defined and employed it. In his pioneering studies of the adaptive problem of visual perception, David Marr started from the premise that the cognitive mechanisms responsible for vision evolved and acquired their "design features" as solutions to particular adaptive problems. Arguing that the best way to understand these mechanisms was to first understand the nature of the problems they were "designed" to solve, he maintained that a "computational theory" of an information-processing problem must be developed before progress can be made in experimentally investigating the cognitive programs that solve it (e.g., Marr, 1982; Marr & Nishihara, 1978). A computational theory specifies the nature of an information-processing problem. It does this by incorporating "constraints on the way the world is structured—constraints that provide sufficient information to allow the processing to succeed" (Marr & Nishihara, 1978, p. 41). A computational theory is an answer to the question: what must happen if a particular function is to be accomplished?[1]

[1]The term "computational theory" is used in two different ways in cognitive psychology: (1) it is used in Marr's sense, as described here, and (2) it is used to refer to a fully specified set of information-processing procedures, usually as realized in a computer simulation. Marr argued that a computational theory in his sense should be developed before a computational theory in this second sense is attempted. In this article, I use the term only in Marr's sense.

Natural selection, in a particular ecological situation, constrains which kinds of traits can evolve. For many domains of human activity, evolutionary biology can be used to determine what kind of psychological mechanisms would have been quickly selected out, and what kind were likely to have become universal and species-typical. Natural selection therefore constitutes "valid constraints on the way the world is structured"; hence, knowledge of natural selection can be used to create computational theories of adaptive information-processing problems. Natural selection theory allows one to pinpoint adaptive problems that the human mind must be able to solve with special efficiency, and it suggests design features that any mechanism capable of solving these problems must have. Of equal importance, evolutionary biology provides the definition of "successful processing" that is most relevant to the study of *biological* information-processing systems: it gives technical content to the concept of "function" (Dawkins, 1982; Williams, 1966), telling the psychologist what adaptive goals our cognitive mechanisms must be able to accomplish.

The approach employed by Marr and others—developing computational theories of a problem defined in functional terms—has been very successful, especially in the field of perception, where the function or goal of "successful processing" is intuitively obvious (i.e., construction of an accurate representation of the three-dimensional world of objects). However, for most kinds of adaptive problem (and, therefore, for most of our cognitive mechanisms), "function" is far from obvious, and intuition uninformed by modern biology is unreliable or misleading. In social cognition, for example, what constitutes adaptive or functional reasoning is a sophisticated biological problem in itself, and is not susceptible to impressionistic, ad hoc theorizing. There exists no domain-general standard for adaptation or "successful processing", therefore functionality must be assessed through reference to evolutionary theory, adaptive problem by adaptive problem.

Fortunately, over the last twenty years, modern evolutionary biology has experienced rapid advances in the technical theory of adaptation, furnishing a series of sophisticated models of what constitutes adaptive behavior in different domains of human activity (e.g., Dawkins, 1982; Hamilton, 1964; Maynard Smith, 1982; Tooby & DeVore, 1987; Trivers, 1971, 1972, 1974; Williams, 1966). Since adaptive behavior is predicated on adaptive thought, these models define a series of adaptive information-processing problems that the human mind must be able to solve with special efficiency. It is therefore possible to develop, out of particular areas of evolutionary biology, computational theories of the specialized cognitive abilities necessary for adaptive conduct in humans.

Through the computational theory, evolutionary biology allows the match-

ing of algorithm to adaptive problem: natural selection theory defines information-processing problems that the mind must be able to solve, and the task of cognitive psychology is to uncover the nature of the algorithms that solve them. Evolutionary biology can be very useful to cognitive scientists if it is used to construct a series of computational models corresponding to the various domains of adaptation that emerged during the course of human evolution; the theory and experiments reported herein are offered as a small illustration of its potential.[2]

Domain-general versus domain-specific explanations of the content effect on the Wason selection task

Although an evolutionary–functional viewpoint, at least in broad outline, should be generally uncontroversial, one implication of such a perspective is not yet widely accepted: that the innate cognitive architecture of the human mind does not simply consist of a few powerful domain-general mechanisms, as many suppose, but instead contains a large array of special-purpose mechanisms, designed to solve an array of recurrent, highly specialized adaptive problems (e.g., Chomsky, 1975, 1980; Cosmides, 1985; Cosmides & Tooby, 1987; Fodor, 1983; Rozin, 1976; Rozin & Schull, 1988; Shepard, 1981; Staddon, 1987; Symons, 1987; Tooby, 1985). There are many reasons to suspect that a small number of domain-general mechanisms can be shown to be inadequate *in principle* to account for adaptive behavior (Cosmides & Tooby, 1987; Symons, 1987); however, this debate will ultimately be settled by the weight of accumulated experimental evidence.

The series of experiments reported herein are meant to contribute to this debate, at least in the area of human reasoning studies. Specifically, an evolutionary perspective suggests that natural selection has shaped how humans reason by creating specialized, domain-specific cognitive mechanisms "designed" to solve discrete adaptive problems by activating reasoning procedures appropriate to the domain encountered. *Evidence for the existence of such mechanisms is: (1) reasoning performance is altered depending on what content the subject is asked to reason about; and (2) such reasoning perfor-*

[2]Note that the method of evolutionary psychology outlined here (and especially in Cosmides & Tooby, 1987) is hypothetico-deductive, rather than speculative. In a speculative approach, one first discovers a psychological mechanism, and then one speculates about what adaptive problem it evolved to solve. The approach advocated here is the reverse: first, one uses existing and validated theories from evolutionary biology to define an adaptive problem that the human mind must be able to solve, and to deduce what properties a psychological mechanism capable of solving that problem must have. Then one tests to see whether there is evidence for a psychological mechanism having the hypothesized properties. It is a constrained and predictive approach, rather than a compilation of post hoc explanations for known phenomena.

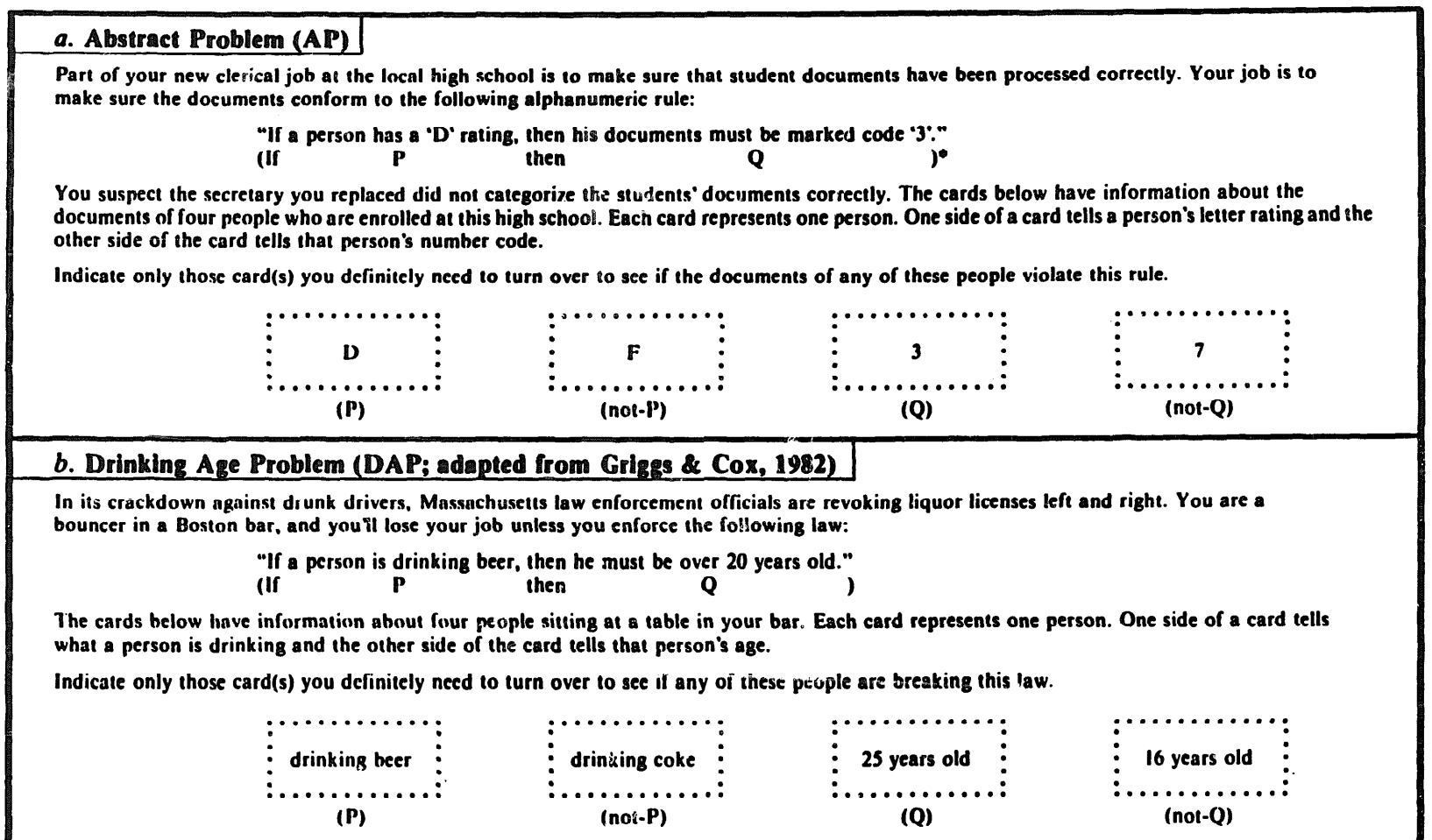
mance is altered by specific content in the predicted adaptive direction. Thus, theories for performance on reasoning tasks can be contrasted by whether the cognitive processes postulated are domain-specific or domain-general, by whether they predict content-dependent or content-independent performance, and by what kinds of content-dependent performance they predict.

The study of human reasoning has been dominated by the search for content-independent cognitive processes. Early research started from the premise that humans reason logically, that is, using the rules of inference of the propositional calculus (e.g., Henle, 1962; Inhelder & Piaget, 1958; Johnson-Laird, 1982; Wason & Johnson-Laird, 1972). These rules of inference are content-independent: they generate only true conclusions from true premises, regardless of what the propositional content of the premises is.

However, more than a decade of research has shown that people rarely reason according to these canons of formal logic (Evans & Lynch, 1973; Johnson-Laird, 1982; Pollard, 1982; Wason & Johnson-Laird, 1972). Moreover—and contrary to initial expectations—psychologists found that human reasoning is content-dependent: the subject matter one is asked to reason about seems to regulate how people reason. Nowhere is this seen more clearly than in experiments using the Wason selection task (Wason, 1966), a test of logical reasoning in which one is asked to determine whether a conditional rule has been violated (see Figure 1). The content of some rules elicits a high percentage of logical responses, whereas the content of other rules does not (Bracewell & Hidi, 1974; Brown, Keats, Keats, & Seggie, 1980; Cox & Griggs, 1982; Gilhooly & Falconer, 1974; Golding, 1981; Griggs & Cox, 1982, 1983; Johnson-Laird, Legrenzi, & Legrenzi, 1972; Manktelow & Evans, 1979; Pollard, 1981; D'Andrade, in Rumelhart & Norman, 1981; Van Duyne, 1974; Wason & Shapiro, 1971; Yachanin & Tweney, 1982). This effect is known as the "content effect" on the Wason selection task.

The discovery of content effects on the Wason selection task did not cause researchers to abandon the hypothesis that human reasoning is governed by content-independent cognitive processes (e.g., Griggs & Cox, 1982; Johnson-Laird, 1982; Manktelow & Evans, 1979; Pollard, 1982; Wason, 1983). By positing that subjects had different amounts of experience with the various content domains tested, experimenters tried to invoke content-independent processes—associationism and an availability heuristic—to explain performance that was manifestly content-dependent. Unfortunately, none of these theories has met with clear predictive success (for review, see Cosmides, 1985). More recently, Cheng and Holyoak (1985) have tried to resurrect this argument by pushing the same explanatory variables one step back, proposing that humans reason using "pragmatic reasoning schemas" that were induced through recurrent experience within goal-defined domains. On this view, the

Figure 1. *Content effects on the Wason selection task. The logical structures of these two Wason selection tasks are identical; they differ only in propositional content. Regardless of content, the logical solution to both problems is the same:* P & not-Q. *Although only 4–25% of college students choose both these cards for the abstract problem* (a), *75% do for the drinking-age problem* (b)*—a familiar "standard social contract".*



*The logical categories (*P*s and *Q*s) marked on the rules and cards are here only for the reader's benefit; they never appear on problems given to subjects.

schemas themselves are content-dependent, but they were created by inductive cognitive processes that are content-independent, and differential experience is a primary variable explaining which schemas are built and which are not. However, Cheng and Holyoak's approach also has many theoretical and empirical problems, as will be discussed below.

An evolutionary analysis suggests that these explanations fail because their basic assumption—that the same cognitive processes govern reasoning about different domains—is false. The more important the adaptive problem, the more intensely selection should have specialized and improved the performance of the mechanism for solving it (Darwin, 1859/1958; Williams, 1966). Thus, the realization that the human mind evolved to accomplish adaptive ends indicates that natural selection would have produced special-purpose, domain-specific, mental algorithms—including rules of inference—for solving important and recurrent adaptive problems (such as learning a language; Chomsky, 1975, 1980). It is advantageous to reason adaptively, instead of logically, when this allows one to draw conclusions that are likely to be true, but cannot be inferred by strict adherence to the propositional calculus. Adaptive algorithms would be selected to contain expectations about specific domains that have proven reliable over a species' evolutionary history. These expectations would differ from domain to domain. Consequently, if natural selection had shaped how humans reason, reasoning about different domains would be governed by different, content-dependent, cognitive processes (Cosmides, 1985; Cosmides & Tooby, 1987, 1989; Rozin, 1976; Rozin & Schull, 1988; Symons, 1987; Tooby, 1985).

Cosmides (1985) and Cosmides and Tooby (1989) used evolutionary theory to develop a computational theory of social exchange—adaptive cooperation between two or more individuals for mutual benefit—which is an evolutionary problem crucial to human adaptation. This computational theory, and its derived implications about the structure of the mental algorithms regulating reasoning about this domain, will be termed *social contract theory*. Because adaptive inference in this domain sometimes converges—and sometimes diverges—from formal logic, it is possible to see if the pattern of reasoning predicted by the proposed social contract algorithms can account for the unexplained content effects that people display in logical reasoning tests. This article reports a series of empirical investigations, using the Wason selection task, of the hypothesis that humans have algorithms specialized for reasoning about social exchange. Three sets of theories are contrasted in these studies: (1) the associationism-based *availability theories* of reasoning; (2) Cheng and Holyoak's induction-based *pragmatic reasoning theory*; and (3) evolutionarily-based *social contract theory*. Consequently, this article is divided into two parts. Part I reports experiments that critically test between

social contract theory and the availability theories of reasoning; Part II reports experiments that critically test between social contract theory and pragmatic reasoning theory's hypothesis that people have a "permission schema". I argue that:

(1) Predictions derived from social contract theory account for existing results better and more parsimoniously than either of the alternative theories.
(2) The experiments reported herein, which use Wason selection tasks of varying content to critically test among these theories, indicate that social contract theory predicts subject performance successfully, whereas the other two theories do not.
(3) Finally, although the ontogenetic origin of specialized reasoning algorithms is not tested by these experiments, their results, taken in the context of the array of results established in the reasoning literature, can be used to assess the *plausibility* of hypotheses that invoke domain-general processes in order to explain the development of the domain-specific procedures that appear to exist.

Darwinian algorithms as adaptive information-processing procedures

When modern evolutionary biology is used to construct computational theories of adaptive information-processing problems, one lesson quickly becomes clear: *although some mechanisms in the cognitive architecture are surely domain-general, these could not have produced fit behavior under Pleistocene era conditions*[3] *(and therefore could not have been selected for) unless they were embedded in a constellation of specialized, content-dependent mechanisms* (Cosmides, 1985; Cosmides & Tooby, 1987; Symons, 1987; Tooby, 1985). Recent evolutionary analyses have shown that there are many domains of human activity for which the evolutionarily appropriate information-processing strategy is extremely complex, and deviations from this strategy result in large fitness costs (e.g., Axelrod & Hamilton, 1981; Hamilton, 1964; Maynard Smith, 1982; Trivers, 1971, 1972, 1974). Only a very

[3]Our species spent over 99% of its evolutionary history as Pleistocene hunter–gatherers: the genus *Homo* emerged about 2 million years ago, and agriculture first appeared less than 10,000 years ago (Lee & DeVore, 1968). Ten thousand years is not long enough for much evolutionary change to have occurred, given the long human generation time; thus, our cognitive mechanisms should be adapted to the hunter–gatherer mode of life, and not to the twentieth century industrialized world. Also, for technical definitions of fitness, see Dawkins, 1982.

narrow subset of behaviors are adaptive in these domains; even supposing there were a domain-independent definition of error, an individual who relied on the vagaries of a domain-general form of induction or trial and error learning would be at a severe selective disadvantage (Cosmides, 1985; Cosmides & Tooby, 1987; Rozin, 1976; Rozin & Schull, 1988; Shepard, 1987; Symons, 1987; Tooby, 1985). Behavioral "plasticity" or "flexibility" is evolutionary death, unless it is accompanied by information-processing mechanisms that are specialized enough to guide behavior into the narrow envelope of adaptive conduct.

The organism's behavior will be random with respect to the constraints that adaptive problems impose unless: (1) it has some reliable and efficient means of extracting information relevant to solving these problems from its environment; and (2) it has well-defined decision rules that use this information in ways that satisfy the constraints. A cognitive system can generate adaptive behavior only if it can perform the specific information-processing tasks entailed by the need to satisfy these constraints.

Consequently, for evolutionarily important problem domains, humans must have evolved "Darwinian algorithms"—specialized learning mechanisms that organize experience into adaptively meaningful schemas or frames. When activated by appropriate problem content, these innately specified "frame-builders" should focus attention, organize perception and memory, and call up specialized procedural knowledge that will lead to domain-appropriate inferences, judgments and choices. Like Chomsky's language acquisition device, these inferential procedures allow one to "go beyond the information given" (Bruner, 1973)—to reason adaptively even in the face of incomplete or degraded information.

Darwinian algorithms regulating social exchange: What properties should they have?

To test the productiveness of this view for cognition, I chose an extensively analyzed area in evolutionary biology: the rules governing the formation of, adherence to, and violation of, social contracts (Axelrod, 1984; Axelrod & Hamilton, 1981; Trivers, 1971). Social exchange—cooperation between two or more individuals for mutual benefit—is biologically rare: few of the many species on earth have evolved the specialized capacities necessary to engage in it (Axelrod & Hamilton, 1981). Humans, however, are one of these species, and social exchange is a pervasive aspect of all human cultures. The ecological and life-historical conditions necessary for the evolution of social exchange were manifest during hominid evolution. Pleistocene small-group living and the advantages of cooperation in hunting and gathering afforded

many opportunities for individuals to increase their fitness through the exchange of goods, services and privileges over the course of a lifetime (Isaac, 1978; Tooby & DeVore, 1987; Trivers, 1971).

Evolutionary biology places tight constraints on how humans must process information regarding social exchange (Axelrod, 1984; Axelrod & Hamilton, 1981; Trivers, 1971). Cosmides (1985) and Cosmides and Tooby (1989) used these constraints to develop a computational theory of social exchange, and found that any algorithm capable of solving this adaptive problem must have certain specific design features, including:

(1) *The human mind must contain algorithms that produce and operate on cost-benefit representations of exchange interactions.* In an exchange, an individual is usually obliged to pay a cost in order to be entitled to receive a benefit. However, the capacity to engage in social exchange could not have evolved unless, on average, both participants realized a net benefit by engaging in exchange: they needed to be able to avoid exchanges in which the cost exceeded the benefit (Axelrod, 1984; Axelrod & Hamilton, 1981; Trivers, 1971). This required specialized cognitive mechanisms that could assess the costs and benefits of various courses of action, and then operate on this information to decide (among other things) whether the benefits of a potential exchange outweighed its costs.[4]

(2) *The human mind must include inferential procedures that make one very good at detecting cheating on social contracts.* The game-theoretic complexities governing conditions of reciprocation in social exchange indicate that the capacity to engage in social exchange *cannot evolve* in a species unless one is able to detect individuals who cheat (fail to reciprocate) on social contracts (Axelrod, 1984; Axelrod & Hamilton, 1981; Trivers, 1971). An individual who engaged in exchange, but who lacked the ability to detect cheaters, would experience fitness costs with no compensating benefits, and would be selected out.

---

[4]The more limited the range of items exchanged by a species, the more item-specific the algorithms regulating social exchange can (and should) be (Cosmides, 1985, chap. 5; Cosmides & Tooby, 1989). For example, the algorithms regulating cleaning fish symbiosis (Trivers, 1971) can be simple and specific: the cleaner fish need only recognize its host, and upon recognizing it, eat its ectoparasites; the host need only recognize the cleaner fish, and upon recognizing it, refrain from eating it. However, because hominid exchanges involved a wide and ever-changing array of items (including tools, information about tool-making, and participation in opportunistically-created, coordinated behavioral routines) Cosmides, and Cosmides and Tooby have predicted that the algorithms regulating social exchange in humans will be item-independent, and will operate on cost–benefit representations of the exchange interaction. Therefore, social contract algorithms should be able to handle a wide variety of items of exchange, as long as these items are *perceived* as costs and benefits to the individuals involved in the exchange.

Specifically, cheating can be defined as the violation of a rule established, explicitly or implicitly, by acceptance of a social contract. A social contract relates *perceived benefits* to *perceived costs*, expressing an exchange in which an individual is required to pay a cost (or meet a requirement) to an individual (or group) in order to be eligible to receive a benefit from that individual (or group). Cheating is the failure to pay a cost to which one has obligated oneself by accepting a benefit, and without which the other person would not have agreed to provide the benefit (Cosmides, 1985). The algorithms that regulate human social exchange—the "social contract algorithms"—should include a "look for cheaters" procedure. In a social exchange situation for which a subject has incomplete information, a "look for cheaters" procedure would draw attention to any person who has *not* paid the required cost (has he illicitly absconded with the benefit?) and to any person who has accepted the benefit (has he paid the required cost?). Such a procedure, operating on the cost–benefit representation of a social contract, maps directly onto the Wason selection task.

The Wason selection task

The Wason selection task (Wason, 1983; see Figure 1) is a paper and pencil problem that invites a subject to see if a conditional rule of the form *If P then Q* has been violated by any one of four instances about which the subject has incomplete information. Each instance is represented by a card. One side of a card tells whether the antecedent is true or false (i.e., whether *P* or *not-P* is the case), and the other side of that card tells whether the consequent is true or false (i.e., whether *Q* or *not-Q* is the case). The subject, who is permitted to see only one side of each card, is asked to say which card(s) must be turned over to see if any of them violate the rule. The four cards that the subject must choose from display terms representing the values *P*, *not-P*, *Q*, and *not-Q*.

From the point of view of formal logic, only the combination on the same card of a true antecedent (*P*) with a false consequent (*not-Q*) can falsify a conditional rule. Thus, regardless of content, the logically correct response to the Wason selection task is to choose the *P* card (to see if it has a *not-Q* on the other side) and to choose the *not-Q* card (to see if it has a *P* on the other side). The card displaying *not-P* and the card displaying *Q* need not be chosen because any value on the other side is consistent with the rule. Although *P* & *not-Q* is the logically correct response, when the content of the conditional rule tested is "abstract" (relates letters to numbers), few subjects make it (4–10%; Wason, 1983). Most choose *P* alone, or *P* & *Q*. A large and contradictory experimental literature revolves around the question of

whether there is any particular sort of problem content that can reliably elicit logically falsifying (*P & not-Q*) responses to the Wason selection task.

Figure 2 shows the cost–benefit structure of a Wason selection task that instantiates a social contract. Irrespective of logical category, a "look for cheaters" procedure would cause the subject to:

(1) Choose the "cost NOT paid" card and the "benefit accepted" card. These cards represent potential cheaters.
(2) Ignore the "cost paid" card and the "benefit NOT accepted" card. These cards represent people who could not possibly have cheated.

As Figure 2 shows, the logical category to which each card corresponds varies, and is determined by where the costs and benefits to the potential cheater are located in the "If–then" structure of the rule. For a "standard" social

Figure 2. *The cost–benefit structure of a Wason selection task that tests a social contract (SC) rule.*

It is your job to enforce the following law:

Rule 1 -- Standard Social Contract (STD-SC): "If you take the benefit, then you pay the cost."
(If P then Q)
Rule 2 -- Switched Social Contract (SWC-SC): "If you pay the cost, then you take the benefit."
(If P then Q)

The cards below have information about four people. Each card represents one person. One side of a card tells whether a person accepted the benefit, and the other side of the card tells whether that person paid the cost.

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

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

Correct social contract answers versus logically correct answers:

	Social Contract answers		Logical answers	
	P & not-Q	*not-P & Q*	*P & not-Q*	*not-P & Q*
standard-SC:	yes	no	yes	no
switched-SC:	no	yes	yes	no

contract like Rule 1 of Figure 2 ("If you take the benefit, then you pay the cost"), the two chosen cards correspond to the logical categories *not-Q* and *P*, respectively. However, for a "switched" social contract like Rule 2 of Figure 2 ("If you pay the cost, then you take the benefit"), the same two cards correspond to the logical categories *not-P* and *Q*.

The correct formal logic response is *P & not-Q*, regardless of content. Therefore, a subject using a "look for cheaters" procedure would appear to be reasoning *logically*—by choosing *P & not-Q*—on standard social contract rules, yet appear to be reasoning *illogically*—by choosing *not-P & Q*—on switched social contract rules.

In addition, the two cards that a "look for cheaters" procedure would ignore correspond to *P & not-Q* (the logically correct response) for a switched social contract rule, and to *not-P & Q* for a standard social contract rule. Hence, social contract theory also predicts that the dominant response to a standard social contract problem will be very rare on a switched one, and vice versa.

Thus, social contract theory makes very specific predictions regarding what kind of logical "errors" subjects will, and will not, make for different kinds of social contract problems. As Figure 2 shows, the correct social contract answers to standard and switched social contract rules differ from the logically correct answers. Therefore, by comparing performance on standard and switched social contract rules, one can tell if reasoning is governed by a logical procedure or by a "look for cheaters" procedure.

Because many competing theories of reasoning performance invoke differential experience as their explanatory variable (e.g., the various availability theories of reasoning), the manipulation of the dimension of familiarity provides an avenue for testing among candidate theories. Because it is proposed that Darwinian algorithms function, in part, as frame *builders* that structure new experiences, social contract theory predicts that the social contract algorithms will operate in unfamiliar situations. No matter how unfamiliar the relation or terms of a rule, if the subject perceives the terms as representing a rationed benefit and a cost requirement—i.e., if the subject recognizes the situation as one of social exchange—then the "look for cheaters" procedure should produce the above pattern of responses. Non-social contract rules, either descriptive or prescriptive, should not show this particular pattern of variation, regardless of their familiarity. In general, they can be expected to elicit the same low levels of *P & not-Q* and very low levels of *not-P & Q* typically found in the literature for non-social contract problems.

Previous results on the Wason selection task are consistent with a social contract interpretation (for a detailed review, see Cosmides, 1985). Robust and replicable content effects are found only for rules that relate terms that

are recognizable as benefits and costs in the format of a standard social contract. No thematic rule that is not a social contract (e.g., rules about food, transportation or school) has ever produced a content effect that is both robust and replicable. For thematic content areas that do not express social contracts, either no content effect is found (e.g., food problems), or there are at least as many studies that do not find content effects as there are studies that do (e.g., transportation and school problems). Moreover, most of the content effects reported for non-social contract rules are either weak (Gilhooly & Falconer, 1974; Pollard, 1981), clouded by procedural difficulties (Bracewell & Hidi, 1974; Van Duyne, 1974), or have some earmarks of a social contract problem (Van Duyne, 1974). All told, for non-social contract thematic problems, 3 experiments have produced a substantial content effect (transportation: Bracewell & Hidi, 1974; Wason & Shapiro, 1971; school: Van Duyne, 1974), 2 have produced a weak content effect (transportation: Gilhooly & Falconer, 1974; Pollard, 1981), and 14 have produced no content effect at all (transportation: Bracewell & Hidi, 1974; Brown et al., 1980; Griggs & Cox, 1982; Manktelow & Evans, 1979; Yachanin & Tweney, 1982; food: Manktelow & Evans, 1979 (4 experiments); Reich & Ruth, 1982; Yachanin & Tweney, 1982; school: Yachanin & Tweney, 1982; non-social contract post office: Golding, 1981; Griggs & Cox, 1982). The few effects that were found did not replicate. In contrast, 16 out of 16 experiments with standard social contracts elicited substantial content effects (Cox & Griggs, 1982; Golding, 1981; Griggs & Cox, 1982, 1983 (10 replications); Johnson-Laird, Legrenzi, & Legrenzi, 1972; D'Andrade, in Rumelhart & Norman, 1981; Van Duyne, 1976). Moreover, deformed social contracts—rules that share constituents with proper social contracts but grossly violate the principles of social exchange (due to their cost–benefit and implicational structures, see Part II)—do not elicit content effects (Griggs & Cox, 1983; for analysis, see Cosmides, 1985, pp. 60–69). In this extensive literature, standard social contract rules are the only thematic rules to elicit strong and replicable content effects on the Wason selection task.

However, none of these studies tested switched social contract rules—rules for which the correct social contract answer is *not-P & Q*. Moreover, most of them contrasted *familiar* standard social contract rules with *unfamiliar* non-social contract rules (descriptive non-social contract rules: Cox & Griggs, 1982; Griggs & Cox, 1982, 1983; Johnson-Laird et al., 1972; prescriptive non-social contract rules: Cox & Griggs, 1982; Golding, 1981; D'Andrade, in Rumelhart & Norman, 1981). Hence, these studies do not allow one to choose directly between a social contract explanation and the explanation that has been most prevalent in the literature, "availability."

PART I: TESTING THE AVAILABILITY THEORIES OF REASONING

The majority of theories proposed in the literature try to account for content effects on the Wason selection task by suggesting that the various subjects tested have had different amounts of experience with the different content domains tested (e.g., Griggs & Cox, 1982; Johnson-Laird, 1982; Manktelow & Evans, 1979; Pollard, 1982; Wason, 1983). Most wed associationism to Tversky and Kahneman's (1973) "availability heuristic", thereby invoking content-independent cognitive processes to explain performance that is content-dependent. Associationism is a process that makes unfamiliar content domains familiar—regardless of the specific content of the domain it operates upon. Which content domains become familiar is determined by the amount of personal experience a particular individual has with the various domains in question. It is a content-independent cognitive process: its features do not differ from domain to domain. The same is true of the availability heuristic: although the processes that make some information more available than others may be content-dependent, the heuristic itself is not.

These "availability" theories come in a variety of forms with some important theoretical differences, but common to all is the notion that the subject's actual past experiences create associational links between terms mentioned in the selection task. The more exposures a subject has had to, for example, the co-occurrence of *P* and *Q*, the stronger that association will be, and the easier it will come to mind—become "available" as a response. A subject is more likely to have actually experienced the co-occurrence of *P & not-Q* for a familiar rule, therefore familiar rules are more likely to elicit logically falsifying responses than unfamiliar rules. However, if all the terms in a task are *unfamiliar*, the only associational link available will be that created between *P* and *Q* by the conditional rule itself, because no previous link will exist among any of the terms. Thus *P & Q* will be the most common response for unfamiliar rules. Falsifying responses will be rare for all unfamiliar rules, whether they are social contracts or not.

For example, the following is Pollard's (1982) summary of his differential availability hypothesis:

> If a selection task uses a realistic P-Q rule, then more Q, or more not-Q, selections will tend to be made, dependent upon whether P-Q or P-not-Q, respectively, is the more available link. Having selected P, the subject will tend to select the card that does, and reject the card that does not, complete the more available link. (p. 83)

And Griggs and Cox (1982) state that their

> ... memory-cueing hypothesis ... proposes performance on the selection task is significantly facilitated when the presentation of the task allows the subject to recall past experience with the content of the problem, the relationship expressed, and a counter-example to the rule governing the relationship. (p. 417)

There is a good reason why these researchers focus so strongly on subjects' actual past experiences: otherwise, they cannot explain why some thematic rules never elicit a content effect (e.g., food problem), and why other thematic rules sometimes elicit an effect and sometimes not (e.g., transportation problem). For example, Pollard (1981) says:

> The hypothesis that the effectiveness of thematic content depends on its relation to the subjects' experience may also be used as a possible explanation of why certain contents fail to produce an effect. For instance, for the non-effective materials reported by Bracewell and Hidi, it is plausible to presume that subjects had no experience of, for instance, "thinking of Ottwa" being related to "remembering car". Counter-examples, therefore, are unlikely to "come to mind" and, as for abstract material, the only cues are those contained in the rule itself. A similar argument applies to the "food and drinks" content of Manktelow and Evans ... it is unlikely that, on the basis of most subjects' experience, the drinking of champagne (or, for that matter, gin) while eating haddock would "come to mind". (p. 27)

Griggs and Cox (1982) suggest that the following might explain why the transportation problem (a rule relating cities and transport, such as "If I go to Boston, then I take the subway") sometimes elicited an effect, and sometimes did not:

> In the Wason and Shapiro study the University of London students might well be familiar with the disconfirming instances (going to Leeds and Manchester by car or train) through their past experience. However, the Plymouth Polytechnic subjects in the Manktelow and Evans (1979) and Pollard (1981) experiments might be less familiar with such journeys since they would be farther away from the cities involved and probably less likely to have actually made the journeys. Such speculation could explain why the Plymouth subjects would show a much reduced effect in the Pollard study or no effect in the Manktelow and Evans study. (p. 419)

Pollard (1982) concurs:

> ... the extent of bias toward one mode of transport would be expected to vary from study to study and, to some extent, from subject to subject, depending on such factors as geographical location, income level of the subjects and the appearance of the experimenter himself (subjects, for instance, may well have experience of professors, but not of postgraduate students, reporting travel by plane). (pp. 80–81)

Thus, when testing against this family of theories, it is important to use rules and terms that are very unfamiliar to one's subjects.

Cheng and Holyoak's tests of the availability theories

In a recent proposal (tested against social contract theory in Part II of this article), Cheng and Holyoak (1985) challenged the availability theories of reasoning. Although this was not their intention, their experiments pitted social contract rules against (1) prescriptive "permission" rules that did not have the clear cost–benefit structure of a social contract, and (2) a rule relating letters to numbers. They found superior performance for the social contract rules, which is the result that social contract theory would predict. However, their experimental design contains a number of confounds that prevent one from ruling out availability as an explanation of their results.

In their first experiment, Cheng and Holyoak used the Wason selection task to investigate prescriptive "permission" rules that were unfamiliar to their subjects (postal rule: "If a letter is sealed, then it must carry a 20-cent stamp"; immigration office rule: "If the form says 'ENTERING' on one side, then the other side includes cholera among the list of diseases"). According to the availability theories of reasoning, unfamiliar rules should not elicit a content effect. However, Cheng and Holyoak believed that if such rules were given a "social purpose", then they would elicit high levels of *P & not-Q* responses (the reasons why are discussed at length in Part II). Therefore, each subject was tested on one rule that had been stripped of any context whatsoever, and one rule that had been given a "social purpose" through the addition of contextual information. (As will be discussed in Part II, this contextual information gave the rules a clear cost–benefit structure, thereby making them social contracts.) The rules that had a context elicited significantly more *P & not-Q* responses than the rules that did not (85–90% vs. 55–60%). This result is consistent with social contract theory, and with Cheng and Holyoak's permission schema theory. Unfortunately, it is also consistent with availability theory.

The "memory-cueing" theories associated with Griggs and Cox (1982; Cox & Griggs, 1982) and Manktelow and Evans (1979), in addition to Pollard's (1982) "differential availability" theory, maintain that contextual information can cue relevant or related experiences from long-term memory, making it more probable that a falsifying response will become available, either directly or through "reasoning by analogy."[5] Because Cheng and Holyoak's experi-

[5]Regarding the reasoning by analogy provision, one could counter-argue that in the absence of a theory specifying which conditions permit reasoning by analogy and which do not, this provision renders the avail-

ment compared a rule that had a context with one that did not, the availability theorist could argue that it confirms, rather than falsifies, availability theory. The context provided would have cued many memories for the "social purpose" problem that would have lain dormant for the no-context problem, increasing the probability that a falsifying response would have become available. This difference in memory-cueing could, therefore, have caused the observed difference in performance. To rule out such an explanation, either both rules should have a context, or neither should.

Second, although subjects had, presumably, never encountered the actual rules used, they were not wholly unfamiliar either. Subjects may not have been familiar with paying more postage for the privilege of *sealing* an envelope, however they would have been familiar with regulations requiring one to pay more or less postage depending on a letter's weight, class, destination, and so on. Similarly, most people know that immigration offices are in the business of deciding who may and may not enter a country, and that many countries forbid entry to people who have not been inoculated against certain diseases. Indeed, even the "contextless" rules elicited a falsifying response from 55–60% of subjects tested. This is usually high enough to be considered a content effect in itself; only 19% of subjects in their second experiment falsified on an abstract problem, which is the usual standard for assessing content effects. Cheng and Holyoak themselves admit that "it may be the case that our subjects were sometimes able to provide their own implicit rationales for the stated rules even when none were provided by the experimenter." Thus, the availability theorist could characterize this experiment as one in which a problem that cued many memories was compared with one that cued a moderate number, and note that the number of falsifying responses was directly proportional to the number of memories cued, just as availability theory predicts. This confound is intrinsic to testing subjects with rules containing familiar elements, but can be controlled for, as discussed below, by using extremely unfamiliar terms and relations.

The above problem is exacerbated by the fact that Cheng and Holyoak allowed subjects to flip back and forth between the no-context and social contract problems, changing their answers. This means that a correct answer

ability theories that have it unfalsifiable, thus robbing them of any empirical content (see Cosmides, 1985, pp. 87–89). However, because a context can be seen as a source of retrieval cues, the point that the problem with a context could have cued more experiences that are *directly* relevant would still stand: on virtually any theory of memory organization, the more retrieval cues provided, the more memories should be activated. For example: subjects may have never paid 20 cents postage *in order to* seal a letter, however, before first class rates last went up, many subjects would have sealed letters having less than 20 cents postage, that is, experienced *P & not-Q*. The context mentioning the post office's goal of increasing profits could have reminded subjects of the last time the post office tried to achieve the same goal by raising first class rates, and thereby cued the falsifying association.

on the social contract problem could have transferred to the no-context problem. Cox and Griggs (1982) have shown that such transfer can take place (see Part II). Moreover, Griggs and Cox (1982) have shown that in the *absence* of this confound, the same contextless postal rule elicits no content effect whatsoever. One therefore does not know to what extent experiences cued by the social contract problem influenced reasoning on the no-context problem. Granted, in the absence of a theory of analogy grounded in social contract theory, the availability theories would have to stretch almost beyond recognition to claim that an immigration problem could cue experiences relevant to solving a postal problem (and vice versa). Nevertheless, an experiment that lacks this confound would provide a much cleaner test.

In a second experiment Cheng and Holyoak tried to remedy some of these problems. In order to avoid cueing specific memories through the use of overly familiar contexts or relations, they compared a schematic permission rule that lacked any specific content to a "concrete" rule relating letters to numbers. The "permission" problem read:

> Suppose you are an authority checking whether or not people are obeying certain regulations. The regulations all have the general form, "If one is to take action 'A', then one must first satisfy precondition 'P'." *In other words, in order to be permitted to do "A", one must first have fulfilled prerequisite "P".* (emphasis added)

The "concrete" problem was a standard abstract selection task with the rule, "If a card has an 'A' on one side, then it must have a '4' on the other side." They consider this problem "concrete" because cards with letters and numbers on them are specific entities. This time subjects were not allowed to flip back and forth, changing answers, and the experimenters analyzed data from just the first problem, to avoid any confounds from the transfer effects that clouded the interpretation of their first experiment. Sixty-one per cent of subjects chose *P & not-Q* for the permission problem, compared to only 19% for the concrete problem.

Unfortunately, this experiment suffers from the same context problem as the first experiment does, though to a lesser degree. The concrete problem is given no context, but the permission problem refers to "an authority checking regulations". An even greater difficulty results from the fact that the permission problem includes a paraphrase of the permission rule (see italicized line). The subject receives no such help with the concrete problem.

The most serious problem with this experiment, however, is that although it was intended to eliminate the memory-cueing problem, it actually exacerbated it. By putting the permission rule in its content-free form, subjects are able to draw on a vast well of experiences with its terms. After all, availability

theory makes no claims about whether the terms of the rule need be basic level objects or superordinate categories for a content effect to occur; there is nothing to prevent one from associating entities like "actions to be taken" with "preconditions to be met". Each of us has had a myriad of experiences in which we were supposed to satisfy a precondition before being permitted to take some action. Furthermore, most of us can probably recall at least one instance in which we violated such a rule by taking the action without having first fulfilled the precondition. In contrast, precisely because the other problem is "concrete", the subject has no comparable well of experiences. How many experiences has the typical subject had with rules relating letters and numbers on the opposite sides of cards? Thus, this result is fully explicable in terms of availability theory. In order to eliminate this problem, Cheng and Holyoak would have to compare their permission problem to a non-permission problem like this:

> Suppose you are a scientist checking to see whether certain rules are true. The rules all have the general form, "If one takes action A, then situation B will occur." In other words, situation B always occurs after one takes action A.

This problem mirrors the permission problem in all respects that are relevant to availability theory: the rule is free of specific content, the text contains a rephrasing and a minimal context, and the rule taps a vast well of experiences (everyone has had experiences in which their actions have had consequences), just like the permission rule does. Availability theory could not explain a wide discrepancy in falsifying responses between the permission problem and this one. Unfortunately, the comparison problem that Cheng and Holyoak used had none of these properties.

For these reasons, Cheng and Holyoak's experiments cannot be used to reject the availability theories of reasoning in favor of either social contract theory or their own permission schema theory; their experimental design contains too many confounds. Tests that control for these confounds are needed. The experiments that follow were carefully designed to critically test social contract theory against the availability theories of reasoning. To minimize the possibility that subjects would have previous associations between the terms of the rule, or that relevant experiences would be cued from long-term memory, the experiments in Part I use completely unfamiliar, culturally alien rules, such as "If a man eats cassava root, then he must have a tattoo on his face." In addition, both unfamiliar social contract rules *and* unfamiliar descriptive rules were given a context (having no context for either rule is not an option; a truly unfamiliar rule can be given the cost–benefit structure of a social contract only through the addition of contextual information). The context surrounding the unfamiliar descriptive rules not only gave

meaning to the terms of the rule, but it also suggested an interpretation of the rule in terms of familiar relationships. This allows the strongest possible test of availability theory: if, in spite of all the contextual information surrounding non-social contract problems, they still elicit low levels of *P & not-Q* responses, while unfamiliar standard social contracts elicit high levels, availability theory would be falsified.

Testing social contract theory against availability theory

At present, it is widely believed that some variant of availability theory accounts for all content effects on the Wason selection task. In contrast, social contract theory proposes that for content involving social exchange, a social contract algorithm is the primary regulator of responses. Thus, for social contract theory, the major determinant of responses is whether a rule is a social contract (SC) or descriptive (descriptive rules differ from social contracts along more dimensions than do non-SC prescriptive rules, thus they minimize the possibility of confounds from transfer effects; non-SC prescriptive rules are tested in Part II). For availability theory, the major determinant of responses is whether a rule is familiar or unfamiliar. Because these two variables are orthogonal, one can create an array of four problem types: unfamiliar social contracts, familiar social contracts, unfamiliar descriptive problems, and familiar descriptive problems (see Table 1).

Moreover, there are two kinds of unfamiliar social contract problems: unfamiliar *standard* social contracts and unfamiliar *switched* social contracts. All but the familiar social contract problems, which confound familiarity with being a social contract, can be used to construct critical tests disentangling the following two hypotheses:

Availability hypothesis: Availability is the sole determinant of performance on Wason selection tasks of varying content. This is the null hypothesis from the standpoint of most of the existing literature.

Table 1. *Experimental design*

SC dimension	Availability dimension	
	Familiar	Unfamiliar
Descriptive	Familiar descriptive	Unfamiliar descriptive
Social contract	Familiar social contract	Unfamiliar social contract (standard) (switched)

Social contract hypothesis: Humans have social contract algorithms that are the major determinant of performance on Wason selection tasks whose content involves social exchange.

It is difficult to believe that availability has *no* effect on familiar problems. The social contract hypothesis is silent on this point. Indeed, any effect availability might have in eliciting falsifying responses to familiar descriptive problems can be used as a standard for judging the size of a social contract effect. Social contract algorithms can be said to be a *major* determinant of responses for problems involving social exchange if there are more social contract responses (the "benefit accepted" card and the "cost not paid" card—standard-SC: *P & not-Q*; switched-SC: *not-P & Q*) to *unfamiliar* social contract problems than falsifying responses to *familiar* descriptive problems.

In the experiments that follow, story context was used to transform an unfamiliar rule—such as, "If a man eats cassava root, then he must have a tattoo on his face"—into either an unfamiliar social contract or an unfamiliar descriptive problem. By embedding the same unfamiliar rule in two different stories, one can contextually define that rule either as a social contract or as a descriptive rule. A social contract story contextually defines one term of the unfamiliar rule as a rationed benefit that must be earned and the other term as a cost/requirement. A descriptive story does not define the terms as costs and benefits, but it does contextually link them to familiar concepts and tie them together by a familiar relation (e.g., the context might explain that men with tattoos live in a different location from men without tattoos, and suggest, "Perhaps men are simply eating foods which are most available to them"). In these experiments, the unfamiliar descriptive problems invoke what should be one of the most familiar relations according to standard associationism: spatio-temporal co-occurrence.

Switching the position that an unfamiliar social contract's terms occupy in the "If–then" structure of the rule transforms its theoretical status from standard to switched, or vice versa. Imagine, for example, a story that portrays cassava root as a rationed benefit and having a tattoo as a cost/requirement. Then:

"If a man eats cassava root, then he has a tattoo on his face"
(If a man takes the benefit, then he pays the cost)

has a standard social contract format, whereas:

"If a man has a tattoo on his face, then he eats cassava root"
(If a man pays the cost, then he takes the benefit)

has a switched social contract format. Switching the position of an unfamiliar descriptive problem's terms does not change its theoretical status.

Experiments 1–4 compare performance on unfamiliar social contract problems to performance on unfamiliar and familiar descriptive problems. These experiments permit six critical tests—comparisons for which the social contract hypothesis and the availability hypothesis make radically different predictions. These tests address the following questions:

(1) Does an *unfamiliar* standard social contract elicit the predicted social contract response, *P & not-Q*?
(2) Are there more social contract responses to an *unfamiliar* standard social contract than falsifying responses to a *familiar* descriptive problem?
(3) Does an *unfamiliar* switched social contract elicit the predicted social contract response, *not-P & Q*?
(4) Are there more social contract responses to an *unfamiliar* switched social contract than falsifying responses to a *familiar* descriptive problem?
(5) Is the correct social contract response to a *standard* social contract (*P & not-Q*) very rare for a *switched* social contract?
(6) Is the correct social contract response to a *switched* social contract (*not-P & Q*) very rare for a *standard* social contract?

Experiments 1 and 2

The purpose of Experiments 1 and 2 was to see whether a standard social contract will elicit the predicted social contract response, *P & not-Q*, even when it is unfamiliar. A high percentage of "falsifying" responses to an unfamiliar standard social contract is predicted only by social contract theory. Availability theory predicts a low percentage of falsifying responses to all unfamiliar rules, whether they are social contracts or not. To choose between the two theories, performance on an unfamiliar standard social contract must be compared to performance on an unfamiliar descriptive rule.

Each subject was asked to solve four Wason selection tasks. Theoretically, the problem types can be described as follows:

U-STD-SC: Unfamiliar standard social contract
U-D: Unfamiliar descriptive
AP: Abstract problem
F-D: Familiar descriptive

The abstract problem was a non-social contract (non-SC) prescriptive rule; it was included because abstract problems are commonly used as a standard for assessing availability (Wason, 1983).

The predictions for Experiments 1 and 2 are identical, and are summarized in Table 2, along with the results.

In Experiment 1, the unfamiliar social contracts used expressed laws of a social group. This was because the rules used in the literature on the Wason selection task were invariably expressed as laws. Moreover, when a social contract is expressed as a law, the exact same rule can be used in unfamiliar social contract and unfamiliar descriptive problems.

However, the form of social exchange that evolutionary models of cooperation were initially developed to explain is the private exchange between two consenting individuals. Therefore, social contract algorithms should operate just as well with conditionals that express private exchanges, rules of the form:

"If you do X for me, then I'll do Y for you",

or, equivalently,

"If I do Y for you, then you do X for me".

Experiment 2 tests this prediction. It is identical to Experiment 1, except the unfamiliar social contracts used expressed an exchange between two individuals rather than a social law.

If social contract theory is correct, then the percentage of *P & not-Q* responses for the private exchange of Experiment 2 should be as high as for the social law of Experiment 1. Moreover, because Experiment 2 tests different social contract rules, it serves as a replication of Experiment 1. This is important, given the persistent failures to replicate that have plagued the Wason selection task literature.

Subjects

Forty-eight undergraduates from Harvard University participated in Experiments 1 and 2, 24 in each experiment. All subjects were paid volunteers recruited by advertisement. Experiment 1 had 13 females and 11 males (mean age: 19.4 years); Experiment 2 had 11 females and 13 males (mean age: 20.0 years (no data on age of 3 subjects)).

Materials and procedures

Each subject received a sealed booklet with instructions on the first page, followed by four Wason selection tasks, one per page. Each selection task was embedded in a brief story. Each booklet contained an unfamiliar standard social contract problem, an unfamiliar descriptive problem, a familiar

descriptive problem, and an abstract problem. The order of the four problems was counterbalanced across subjects. All experiments in Part I had a within-subjects design.

The instructions asked subjects to do the four tasks in order, without rereading any previous story or reviewing or changing any previous answers. The instructions were read aloud to subjects; in addition, subjects were given as much time as they wanted to read the instructions over to themselves before breaking the seal and beginning the experiment. Although most subjects completed the experiment in about 10 minutes, they were told they could take as much time as they wanted.

All stories were phrased so as to activate a "detective set" (Van Duyne, 1974), and all asked subjects to look for *violations* of the rule. The unfamiliar rules used and the cultures described in the problems' text were fictitious.

Materials for Experiment 1: The unfamiliar rules (both social contract and descriptive) were: "If a man eats cassava root, then he must have a tattoo on his face", and "If you eat duiker meat, then you have found an ostrich eggshell". To avoid possible transfer effects between the two unfamiliar problems, a booklet contained either the "cassava root" version of the unfamiliar descriptive problem and the "duiker meat" version of the unfamiliar standard social contract, or vice versa. The "cassava root" version of the social contract and descriptive problems varied only in surrounding story context; the rules used were identical. The same was true of the "duiker meat" version. The full text of all problems used is reported in the Appendix.

I wanted to use any effect availability might have in eliciting falsifying responses to the familiar descriptive problem as a standard for judging the size of a social contract effect. For this reason I used a transportation problem as the familiar descriptive problem: the transportation problem had been the most successful non-SC problem in the literature (see Cosmides, 1985). Various versions (using different terms) of the familiar descriptive and abstract problems were randomized with respect to each other and to the unfamiliar problems.

Experiment 2: The procedure for Experiment 2 was identical to that described for Experiment 1. The materials were also identical, with one exception: the unfamiliar standard social contract expressed a private exchange rather than a social law.

Materials for Experiment 2: The unfamiliar social contracts used in Experiment 2 were: "If you get a tattoo on your face, then I'll give you cassava root", and "If you give me your ostrich eggshell, then I'll give you duiker meat". In both cases, the person offering the contract is identified as the potential cheater. Stated in terms of the value system of the potential cheater, the social contract structure of both problems therefore was: "If you benefit

me, then I will pay a cost to you"—a standard social contract. In the stories, the potential cheater always received his benefit *before* he had to pay the required cost.[6] The stories surrounding the private exchanges of Experiment 2 were quite different from those surrounding the social contract laws of Experiment 1 (see Appendix for full text).

The unfamiliar descriptive rules used in Experiment 2 were: "If a man has a tattoo on his face, then he eats cassava root", and "If you have found an ostrich eggshell, then you eat duiker meat". Although no story context can transform a private exchange social contract into a descriptive rule, the terms used in both sets of rules referred to the same kinds of objects: cassava root, facial tattoos, duiker meat and ostrich eggshells. Moreover, the unfamiliar descriptive and social contract problems were matched as to whether the various terms appeared in the antecedent or consequent clause.

Results

Table 2 shows the percentage of subjects choosing *P & not-Q* for each problem in Experiments 1 and 2. The results closely match the social contract predictions: a very high percentage of subjects chose *P & not-Q* in response to the unfamiliar standard social contracts (Exp. 1: 75%; Exp. 2: 71%), but few made this response to the unfamiliar descriptive problems (Exp. 1: 21%; Exp. 2: 25%).

---

[6]Note that both social contract stories include a time delay between when the potential cheater receives his benefit and when he must cough up the required cost—which is the benefit to the other person. In most Pleistocene exchanges reciprocation was delayed, not simultaneous (Cosmides, 1985, chap. 5; Tooby & DeVore, 1987; Trivers, 1971). Cheating is far easier when reciprocation must occur after a benefit has been received, and subjects should be more likely to suspect someone of intending to cheat in such delayed benefit transactions. In a simultaneous, face-to-face exchange, if you see that the other person has come prepared to cheat, you simply withhold what he or she wants. Subjects can be expected to assume that such intercontingent behavior will occur in face-to-face exchanges, unless they are given information to the contrary.

If subjects made this assumption, what would happen to performance on a social contract rule with no time delay? Subjects would fail to choose the "cost NOT paid" card (standard-SC: *not-Q*; switched-SC: *not-P*). This card indicates that the potential cheater had, indeed, come prepared to cheat—that he had NOT paid the cost. The subject would assume that, upon seeing this, the honest party in the interaction would simply withhold the item that the potential cheater had wanted (the benefit to cheater). No exchange would have taken place, and therefore no cheating. Subjects would therefore choose only the "benefit accepted" card: *P* alone on a standard social contract, *Q* alone on a switched one.

Such sophisticated reasoning about intercontingent behavior should come quickly and easily to subjects. In fact, I ran several pilots in which I had forgotten to include a time delay. After doing the tasks, a number of subjects spontaneously told me they had assumed the intercontingent scenario sketched above. Their card choices on both standard and switched social contract problems were consistent with their claim.

Thus, the time delay of "reciprocal altruism" is an essential element in a story when the rule expresses a private exchange: it allows the potential cheater to seize the benefit before he or she is expected to pay the cost. The honest person then has no options.

Table 2. *Experiments 1 and 2, predictions and results: percentage of subjects choosing* P & not-Q *or* not-P & Q *for each problem* (n = *24*)[a]

	P & not-Q responses			
	Predictions		Results	
	Social contract	Availability	Exp. 1	Exp. 2
U-STD-SC:	High	Low	75	71
U-D:	Low	Low	21	25
AP:	Low	Low	25	29
F-D:	Low	Middling to low	46	38

	not-P & Q responses		
	Predictions	Results	
	Both theories[b]	Exp. 1	Exp. 2
U-STD-SC:	Very low	0	0
U-D:	Very low	0	0
AP:	Very low	0	0
F-D:	Very low	0	0

[a]Table 2 compares the results obtained in Experiments 1 and 2 with the predictions of social contract theory and of availability theory (assuming that responses are determined by either SC algorithms or availability, but not both). In Experiment 1 the social contract expressed a law, whereas in Experiment 2 it expressed an exchange between two persons. U-STD-SC = unfamiliar standard social contract; U-D = unfamiliar descriptive problem; AP = abstract problem; F-D = familiar descriptive problem.

[b]N.B.: *not-P & Q* is a very rare response on Wason selection tasks. It involves the failure to choose the *P* card, which is almost universally chosen, and which even availability theorists concede is guided by at least a rudimentary understanding of logic (Evans and Lynch, 1973; Pollard, 1979). In addition, when chosen with *Q*, the substitution of *not-P* for *P* violates ordinary notions of contingency as expressed in English (why say "If P then Q" when one means "If not-P then Q"?). No availability theorist has ever predicted this response. Both hypotheses predict a *very* low percentage for all problems other than a switched social contract, for which, according to social contract theory alone, it is the predicted response.

Residual responses. *P & not-Q* and *not-P & Q* do not exhaust all possible combinations of card choices—there are 16 in all. Experiment 1: Of the 6 responses to the standard social contracts that were not full SC answers, 5 were half correct "sins of omission": 2 *P* responses (omitted *not-Q*) and 3 *not-Q* responses (omitted *P*). Experiment 2: There were 7 responses to the standard social contracts that were not full SC answers, and 5 of these were

half correct "sins of omission"–5 *P* responses. Thus, for the unfamiliar standard social contract problems, 96% of subjects gave responses consistent with social contract theory in Experiment 1, and 92% did in Experiment 2.

The availability and social contract hypotheses both predict that the percentage of subjects choosing *not-P & Q* will be very low for all four problems; indeed, no one made this response for any problem in either experiment.

Critical tests

The critical tests compare problems for which social contract theory and the availability theories make radically different predictions. Predictions for critical tests 1 and 2 are taken from Table 2.

Critical test 1: Does an unfamiliar standard social contract elicit the predicted social contract response, *P & not-Q*?

To answer this question, responses to the two unfamiliar problems must be compared. These problems use the same rule, but the story surrounding one rule makes it a standard social contract (U-STD-SC), whereas the story surrounding the other makes it a descriptive rule (U-D). Because both these rules are unfamiliar, availability theory predicts that both will elicit a low percentage of *P & not-Q* responses. However, social contract theory predicts that a standard social contract will elicit a high percentage of *P & not-Q* responses, even though it is unfamiliar.

Percentage *P & not-Q* responses:

	Social contract prediction:	Availability prediction:
U-STD-SC vs. U-D:	U-STD-SC > U-D	U-STD-SC = U-D
	high low	low low
Exp. 1:	75% > 21%	
Exp. 2:	71% > 25%	

Availability does not predict, and cannot account for, a wide discrepancy in falsifying (*P & not-Q*) responses between these two unfamiliar problems. Yet a highly significant discrepancy occurred in both experiments, just as social contract theory predicts (Exp. 1: 75% vs. 21%: $F(1,23) = 27.18$, $p \ll .001$, $r = .74$. Exp. 2: 71% vs. 25%: $F(1,23) = 19.46$, $p < .001$, $r = .68$). This discrepancy is predicted only by social contract theory; availability theory predicts a low proportion of falsifying responses on all unfamiliar problems, whether they are social contracts or not. The unfamiliar standard social contract also produced a significant "content effect" when measured against the abstract problem (Exp. 1: 75% vs. 25%: $F(1,23) = 13.80$, $p < .005$, $r = .61$. Exp. 2: 71% vs. 29%: $F(1,23) = 16.43$, $p < .001$, $r = .65$). The unfamiliar

descriptive and abstract problems both elicited the same low levels of falsifying responses (Exp. 1: 21% vs. 25%: $F(1,23) = 0.138$, n.s. Exp. 2: 25% vs. 29%: $F(1,23) = 0.19$, n.s.).

Critical test 2: Are there more social contract responses to an *unfamiliar* standard social contract than falsifying responses to a *familiar* descriptive problem?

All problems asked subjects to detect potential violations of the rule. Therefore, any effect availability has in eliciting *falsifying* responses to familiar non-SC problems can be used as a standard for judging the size of the social contract effect.

Percentage *P & not-Q* responses:

	Social contract prediction:		Availability prediction:	
U-STD-SC vs. F-D:	U-STD-SC > F-D		U-STD-SC ≤ F-D	
	high	low[7]	low	mid–low
Exp. 1:	75% > 46%			
Exp. 2:	71% > 38%			

As social contract theory predicts, an unfamiliar social contract (U-STD-SC) with which no subject could have had any actual experience elicited significantly more "falsifying" responses (actually, SC responses) than a familiar relation (F-D) with which subjects were likely to have had experience (Exp. 1: 75% vs. 46%: $F(1,23) = 5.24$, $p < .05$, $r = .43$. Exp. 2: 71% vs. 38%: $F(1,23) = 6.57$, $p < .025$, $r = .47$). Availability predicts an inequality in the opposite direction.

It is interesting to see what happens when one considers residual responses for both problems. *Not-Q* chosen alone is the most interesting residual response because *not-Q* is the card people routinely forget to choose on Wason selection tasks, and it is rarely chosen alone. If one counts this rare residual response for both problems, the unfamiliar standard social contract's advantage is magnified in Experiment 1 (87.5% vs. 54%: $F(1,23) = 5.41$, $p < .05$, $r = .44$), but remains unchanged for Experiment 2.

Experiments 3 and 4

The results of Experiments 1 and 2 support the hypothesis that reasoning on social contract problems is guided by special-purpose social contract al-

[7]The social contract hypothesis is silent on whether availability exerts an independent effect on familiar problems that do not involve social exchange.

gorithms. However, the results are also consistent with an alternative hypothesis: that social contract problems somehow facilitate logical reasoning. This is because the SC answer to a standard social contract is the same as the logically correct answer: *P & not-Q*. To choose between these two hypotheses, one must test social contract rules for which the correct SC answer is *different* from the logically correct answer.

That is the purpose of Experiments 3 and 4. These experiments are identical to Experiments 1 and 2, except the unfamiliar rules were switched rather than standard. Thus, instead of reading "If you take the benefit, then you pay the cost", a switched social contract reads "If you pay the cost, then you take the benefit." Social contract algorithms are, by hypothesis, content-dependent: a "look for cheaters" procedure should always pick the "cost not paid" card and the "benefit accepted" card, regardless of what logical category they happen to fall into. This means that the correct SC answer to a switched social contract is *not-P & Q: not-P* is the "cost not paid" card and *Q* is the "benefit accepted" card.

Logic, in contrast, is content-independent: whether "you take the benefit" occurs in the "if" clause or the "then" clause makes no difference whatsoever from a logical point of view. *P & not-Q* is the *logically* correct answer, regardless of whether the social contract is standard or switched. Note that a logically correct answer to a switched social contract makes no sense from the content-dependent, social contract point of view. It corresponds to choosing the "cost paid" card (*P*) and the "benefit not accepted" card (*not-Q*), which represent two individuals who cannot possibly have cheated; hence, an individual who reasoned logically on a switched social contract would fail to detect potential cheaters.

Therefore, if social contract problems merely facilitate our ability to reason logically, then the unfamiliar *switched* social contracts tested in Experiments 3 and 4 should elicit high percentages of logically falsifying *P & not-Q* responses, just as the unfamiliar *standard* social contracts did in Experiments 1 and 2. In contrast, if social contract algorithms control reasoning about social contract problems, then the switched social contracts should elicit high percentages of *not-P & Q* responses.

In Experiment 3, the social contract rule tested expressed a social law, whereas in Experiment 4 it expressed a private exchange. The four problems tested in each experiment fell into the following theoretical categories:

U-SWC-SC:	Unfamiliar switched social contract
U-D:	Unfamiliar descriptive
AP:	Abstract problem
F-D:	Familiar descriptive

The predictions for Experiments 3 and 4 are identical, and are summarized in Table 3, along with the results.

Subjects

Forty-eight undergraduates from Harvard University participated in Experiments 3 and 4, 24 in each experiment. They were paid volunteers recruited by advertisement, and they had not participated in either of the previous experiments. Both experiments had 11 females and 13 males (mean age: Exp. 3: 19.0 years; Exp. 4: 19.4 years (no data on one subject's age)).

Materials and procedures

The procedure for Experiments 3 and 4 was identical to that for Experiments 1 and 2. The materials were also identical with one exception: for unfamiliar rules, the propositions were switched.

Materials for Experiment 3: This experiment tested an unfamiliar social contract in the form of a social law. It was identical to Experiment 1, except that the unfamiliar rules (both social contract and descriptive) were "switched." Thus the rule tested in the cassava root version of these problems was: "If a man has a tattoo on his face, then he eats cassava root", and the rule tested in the duiker meat version was "If you have found an ostrich eggshell, then you eat duiker meat". Technically, a switched version of the cassava root rule would read "If a man *must* have a tattoo on his face, then he eats cassava root", however, the modal "must" was left out of the "If" clause in the switched version because it violates standard English usage. The standard version of the duiker meat problem did not have a modal "must" to begin with, so no alterations had to be made in switching it.

Materials for Experiment 4: This experiment tested an unfamiliar social contract in the form of a personal exchange. It was identical to Experiment 2, except that the unfamiliar rules (both social contract and descriptive) were "switched." Thus, the switched social contract rules were: "If I give you cassava root, then you must get a tattoo on your face", and "If I give you duiker meat, then you must give me your ostrich eggshell". The person offering the contract was the potential cheater, so from his point of view these rules read: "If I pay a cost to you, then you must benefit me". (The standard form reads: "If you benefit me, then I will pay a cost to you".) To be consistent with common English usage and to keep the time relations of the rule consistent with those of the story (the potential cheater always received his benefit first), in the switched version the word "will" was struck from the antecedent clause and "must" was added to the consequent clause.

The unfamiliar descriptive rules were: "If a man eats cassava root, then he must have a tattoo on his face", and "If you eat duiker meat, then you have found an ostrich eggshell".

Results

Table 3 shows the percentage of subjects choosing *not-P & Q* and *P & not-Q* for each problem: these figures closely match the social contract predictions. If social contract problems merely facilitated logical reasoning, the unfamiliar switched social contract (U-SWC-SC) should have elicited high percentages of falsifying responses. This did not happen: only 1 subject chose *P & not-Q*

Table 3. *Experiments 3 and 4, predictions and results: percentage of subjects choosing* not-P & Q *or* P & not-Q *for each problem* (n = *24*)[a]

	not-P & Q responses			
	Predictions		Results	
	Social contract	Availability	Exp. 3	Exp. 4
U-SWC-SC:	High	Very low[b]	67	75
U-D:	Very low	Very low	4	0
AP:	Very low	Very low	0	0
F-D:	Very low	Very low	0	0

	P & not-Q responses			
	Predictions		Results	
	Social contract	Availability	Exp. 3	Exp. 4
U-SWC-SC:	Very low[c]	Low	4	0
U-D:	Low	Low	12	25
AP:	Low	Low	12	33
F-D:	Low	Middling to low	50	58

[a]Table 3 compares the results obtained in Experiments 3 and 4 with the predictions of social contract theory and of availability theory (assuming that responses are determined by either SC algorithms or availability, but not both). In Experiment 3 the social contract expressed a law, whereas in Experiment 4 it expressed an exchange between two persons. U-SWC-SC = unfamiliar switched social contract; U-D = unfamiliar descriptive problem; AP = abstract problem; F-D = familiar descriptive problem.

[b]For rationale, see notes to Table 2.

[c]*P* and *not-Q* are the cards social contract algorithms should ignore on switched social contract problems; hence the percentage of "falsifying" responses should be *very* low on the U-SWC-SC.

in Experiment 3, and no one chose it in Experiment 4. Moreover, the percentage of subjects who gave the correct social contract answer, *not-P & Q*, was high in both experiments, just as social contract theory predicts (Exp. 3: 67%; Exp. 4: 75%). This indicates that for social contract problems subjects are following the rules of social exchange rather than the rules of formal logic.

Residual responses. Experiment 3: Of the 8 responses that were not full SC answers, 6 were half correct "sins of omission": 1 *not-P* (omitted *Q*) and 5 *Q* (omitted *not-P*). Both are rare answers on Wason selection tasks, and the frequency of *Q* responses is ten times higher than its expected value based on the abstract problems in Experiments 1 and 2 ($Z = 2.71$, $p < .0034$). Experiment 4: Four of the 6 responses that were not full SC answers were half correct "sins of omission": 4 *Q* responses (omitted *not-P*).

Thus, 92% of subjects in both experiments gave answers that were consistent with the social contract theory prediction.

Critical tests

Critical test 3: Does an unfamiliar switched social contract elicit the predicted social contract response, *not-P & Q*?

Adaptive inference diverges sharply from logical inference for switched social contract problems. *Not-P & Q* is completely at variance with formal logic and a very rare response on Wason selection tasks—the availability hypothesis predicts it will be rare for all problems. However, *not-P & Q* is the correct SC response to a switched social contract, no matter how unfamiliar. Therefore, critical test 3 requires that the unfamiliar switched social contract be compared to all the other rules. *Not-P & Q* is the predicted response only for the unfamiliar switched social contract.

Percentage *not-P & Q* responses:

U-SWC-SC vs. U-D, AP, F-D:	Social contract prediction: U-SWC-SC > U-D, AP, F-D		Availability prediction: U-SWC-SC = U-D, AP, F-D	
	high	very low	very low	very low
Exp. 3:	67% >	4%, 0%, 0%		
Exp. 4:	75% >	0%, 0%, 0%		

The large and significant 63–75 point difference between the unfamiliar switched social contract and all other problems is predicted only by social contract theory (Exp. 3: 67% vs. 4%, 0%, 0%, $L = +3, -1, -1, -1$: $F(1,69) = 116.26$, $p \ll .001$, $r = .79$. Exp. 4: 75% vs. 0%, 0%, 0%, $L = +3, -1, -1, -1$: $F(1,69) = 207.01$, $p \ll .001$, $r = .87$). In all four experiments, only

one person chose *not-P & Q* in response to a problem other than a switched social contract.

Critical test 4: Are there more SC responses to an unfamiliar switched social contract than falsifying responses to a familiar descriptive problem?

As in critical test 2, one can use the percentage of falsifying responses to the familiar descriptive problem as a standard for judging the size of the social contract effect. This requires that the proportion of SC responses (*not-P & Q*) to the unfamiliar switched social contract be compared to the proportion of falsifying responses to the familiar descriptive problem.

Percentage *not-P & Q* responses to U-SWC-SC, percentage *P & not-Q* responses to F-D:

	Social contract prediction:		Availability prediction:	
U-SWC-SC vs. F-D:	U-SWC-SC > F-D		U-SWC-SC < F-D	
	high	low	very low	mid-low
Exp. 3:	67% > 50%			
Exp. 4:	75% > 58%			

As the social contract hypothesis predicts, SC responses to the unfamiliar switched social contract outstripped falsifying responses to the familiar descriptive problem in both experiments (Exp. 3: 67% vs. 50%: $F(1,23) = 2.09$, n.s. Exp. 4: 75% vs. 58%: $F(1,23) = 1.64$, n.s.). Although the difference is not significant in either case, the availability hypothesis predicts an inequality in the *opposite* direction.

In both experiments, when rare residuals are counted for both problems (U-SWC-SC: *Q*; F-D: *not-Q*), the difference between the two problems is magnified and is significant (Exp. 3: 87.5% vs. 54%: $F(1,23) = 8.36$, $p < .01$, $r = .52$. Exp. 4: 92% vs. 63%: $F(1,23) = 6.75$, $p < .025$, $r = .48$.

Like the results of critical test 2, which compared unfamiliar *standard* social contracts to familiar descriptive problems, these data support the contention that SC algorithms are a major determinant of responses to problems involving social exchange, even when those problems are unfamiliar.

Critical test 5: Is the correct SC response to a standard social contract (*P & not-Q*) very rare for a switched social contract?

Because the standard and switched social contracts tested were both unfamiliar, the availability hypothesis predicts they should both elicit low levels of *P & not-Q* responses. The social contract prediction could not be more different. For a standard social contract, *P* represents the "benefit accepted" card and *not-Q* represents the "cost NOT paid" card, the cards that a "look

for cheaters" procedure should choose. However, for a switched social contract, *P* represents the "cost paid" card and *not-Q* represents the "benefit NOT accepted" card, the cards that a "look for cheaters" procedure should *ignore* because they represent people who could not possibly have cheated (see Figure 2). Thus *P & not-Q*, the correct SC answer for a standard social contract, should be a very rare response to a switched social contract. One can test this prediction by comparing performance on matched standard and switched social contract rules: the two social laws of Experiments 1 and 3 and the two private exchanges of Experiments 2 and 4.

Percentage *P & not-Q* responses:

	Social contract prediction:	Availability prediction:
U-STD-SC vs. U-SWC-SC:	U-STD-SC ≫ U-SWC-SC	U-STD-SC = U-SWC-SC
	high very low	low low
Exps. 1 and 3:	75% ≫ 4%	
Exps. 2 and 4:	71% ≫ 0%	

The large and significant 71 point discrepancies in *P & not-Q* responses between the standard and switched social contract problems is predicted only by social contract theory (Exps. 1 and 3: 75% vs. 4%: $Z = 5.02, p < .0000005$, $\phi = .72$. Exps. 2 and 4: 71% vs. 0%: $Z = 5.14, p < .00000025, \phi = .74$). Furthermore, the social contract theory prediction that the dominant, SC response to the standard social contract will be very rare on the switched social contract was borne out. In Experiment 3, only one subject gave the standard social contract answer, *P & not-Q*, in response to the switched social contract—and this was one of only two subjects in Experiment 3 to give falsifying answers to all of the three other problems. And no one chose *P & not-Q* in response to the switched social contract in Experiment 4.

Critical test 6: Is the correct SC response to a switched social contract (*not-P & Q*) very rare for a standard social contract?

Critical test 6 is simply the flip side of critical test 5. The predicted SC response to a switched social contract is *not-P & Q*: *not-P* represents the "cost NOT paid" card and *Q* represents the "benefit accepted" card (see Figure 2). But for a standard social contract, *not-P* represents the "benefit NOT accepted" card and *Q* represents the "cost paid" card—the cards that a "look for cheaters" procedure should ignore, regardless of their logical category. Hence, social contract theory predicts that the correct SC answer to a switched social contract, *not-P & Q*, will be very rare for a standard social contract. In contrast, availability predicts that the percentage of subjects choosing *not-P & Q* on the standard and switched social contracts will be about equal, and very low (see Table 2). Again, one must compare

matched laws: the two social laws of Experiments 3 and 1, and the two private exchanges of Experiments 4 and 2.

Percentage *not-P & Q* responses:

	Social contract prediction:		Availability prediction:	
U-SWC-SC vs. U-STD-SC:	U-SWC-SC ≫ U-STD-SC		U-SWC-SC = U-STD-SC	
	high	very low	very low	very low
Exps. 3 and 1:	67% ≫ 0%			
Exps. 4 and 2:	75% ≫ 0%			

The large and significant 67–75 point discrepancy in *not-P & Q* responses between switched and standard social contracts is predicted only by social contract theory (Exps. 3 and 1: 67% vs. 0%: $Z = 4.90$, $p < .0000005$, $\phi = .71$. Exps. 4 and 2: 75% vs. 0%: $Z = 5.37$, $p < .0000001$, $\phi = .77$). Furthermore, the social contract prediction that the dominant, SC response to the switched social contract will be very rare for a standard one was borne out: no one gave the switched social contract answer, *not-P & Q*, in response to the standard social contract in either experiment.

Summary, critical tests for Experiments 1–4

Because the availability and social contract hypotheses make very different predictions regarding six comparisons between problems, critical tests between these two hypotheses can be constructed from the predictions of Tables 2 and 3. Two data sets, each examining a different kind of social contract (social law vs. private exchange), were subjected to each critical test. The results for each of the six critical tests verified the social contract prediction and falsified the availability prediction. This was true for both the social law data and the private exchange data. Thus, results of each critical test replicated.

Social contract tests

Critical tests only address the question: Are the data better explained by social contract theory or by availability theory? However, there are other questions one can ask of this data, questions that are specific to social contract theory.

(1) Are the logically distinct SC answers to standard and switched social contract problems produced by the same algorithms?

According to social contract theory, one should always choose the "benefit accepted" card and the "cost NOT paid" card, regardless of which logical category these fall into. From a logical point of view, these cards fall into very different categories for standard and switched social contracts (standard: *P & not-Q*; switched: *not-P & Q*). If social contract algorithms are sensitive to social contract categories, rather than to logical categories, then standard and switched social contracts should elicit the same percentage of SC answers, even though they are logically distinct.

Indeed, the proportions of SC answers to the standard and switched social contract laws in Experiments 1 and 3 are not significantly different (U-STD-SC = 75%, U-SWC-SC = 67%, $Z = 0.63$, n.s.), just as one would expect if the same SC algorithms were producing these two, logically distinct, responses. When rare residuals are added in, the percentage of SC answers on these two problems is identical—87.5%. Using percentage falsifying answers to the unfamiliar descriptive problems (which had the same rule as unfamiliar social contracts) as a baseline for comparison, the *relative* advantage SC status gave in producing SC answers is almost identical for both SC problems: 54 points between U-STD-SC and its U-D, 55 points between U-SWC-SC and its U-D.

This was also true of the standard and switched personal exchange problems in Experiments 2 and 4 (U-STD-SC = 71%, U-SWC-SC = 75%: $Z = 0.32$, n.s. Relative advantage compared to U-D baseline: U-STD-SC: 46 points; U-SWC-SC: 50 points).

For the personal exchange problems, residual responses did not split (e.g., some *P*, some *not-Q* for the U-STD-SC) as they did for the law problems of Experiments 1 and 3; the only "sins of omission" in Experiments 2 and 4 involved choosing the "benefit accepted" card, that is, the card indicating that the honest person gave the benefit to the potential cheater (5 *P* alone responses for the U-STD-SC, 4 *Q* alone responses to the U-SWC-SC). This is interesting, because only the private exchange experiments admit the possibility of intercontingent behavior: these are the responses one would expect if a subject read through quickly, not noticing that there is a time delay between when the potential cheater gets his benefit and when he is expected to honor his end of the deal (see footnote 6). The numbers involved are too small to firmly attribute this pattern to the power of intercontingent reasoning, but it is a question that deserves further research.

Table 4 shows the frequencies with which individual cards were selected in all four experiments (the results of Exps. 1 and 2, which tested standard social contracts, are collapsed, as are those of Exps. 3 and 4, which tested

Table 4. *Experiments 1–4: selection frequencies for individual cards, sorted by logical category and by social contract category*[a]

	Unfamiliar descriptive		Abstract problem		Familiar descriptive		Unfamiliar social contract	
							Standard	Switched
Logical category	Exps. 1 and 2	Exps. 3 and 4	Exps. 1 and 2	Exps. 3 and 4	Exps. 1 and 2	Exps. 3 and 4	Exps. 1 and 2	Exps. 3 and 4
P	43	40	46	46	46	45	43	3
not-P	9	11	10	11	1	2	3	36
Q	20	23	15	23	8	6	0	44
not-Q	18	20	21	25	23	32	39	3
	Social contract category:		Benefit accepted				43	44
			Benefit NOT accepted				3	3
			Cost paid				0	3
			Cost NOT paid				39	36

[a]Experiments 1 and 2 tested standard versions of the two unfamiliar rules, whereas Experiments 3 and 4 tested switched versions of these rules.

switched ones). When cards are sorted according to their logical category, all problems replicate nicely over the standard and switched experiments, except the social contract problems. When sorted according to logical category, selection frequencies for standard and switched social contracts are radically at variance with one another. When sorted according to social contract category, however, their profiles are almost identical. This indicates that for unfamiliar social contract problems, a social contract categorization scheme captures dimensions that are psychologically real for subjects, whereas a logical categorization scheme does not.

This is what one would expect if the social contract algorithms were activated for both standard and switched social contract problems, but not for non-social contract problems.

(2) How well do SC algorithms operate in novel, versus familiar, social exchanges?

Learning mechanisms should make the unfamiliar familiar. A frame-*builder* is a learning mechanism: it structures new experiences along evolutionarily relevant dimensions that it is keyed to pick up. If SC algorithms are, in part, frame-*builders*, as proposed, one would expect them to operate in novel

social exchanges—like those represented in the unfamiliar social contract problems—as well as in familiar ones.

The drinking-age problem, a highly familiar standard social contract, usually elicits a *P & not-Q* response from 75% of subjects tested (Cox & Griggs, 1982; Griggs & Cox, 1982, 1983). Interestingly enough, just as many *P & not-Q* responses were elicited by the *unfamiliar* standard social contracts tested in Experiments 1 and 2 (Exp. 1: 75%; Exp. 2: 71%). This cannot be accounted for by differences in subject populations: 78% of a similar group of 23 Harvard undergraduates "falsified" on the drinking-age problem (Cosmides, 1985). Thus, the percentage of Harvard undergraduates choosing *P & not-Q* on a standard social contract was the same, regardless of whether the social contract was very familiar or completely unfamiliar (Exp. 1: 75% vs. 78%: $Z = 0.26$, n.s. Exp. 2: 71% vs. 78%: $Z = 0.58$, n.s.).

The same was true of the unfamiliar switched social contract. The percentage of subjects choosing *not-P & Q* on the unfamiliar switched social contracts of Experiments 3 and 4 did not differ significantly from the percentage choosing *P & not-Q* for the familiar drinking-age problem (Exp. 3: 67% vs. 78%: $Z = 0.88$, n.s. Exp. 4: 75% vs. 78%: $Z = 0.26$, n.s.). The hypothesis that unfamiliar social contract problems generate fewer SC answers than familiar ones is not supported even if one uses all three problems (F-STD-SC, U-STD-SC, U-SWC-SC) in one test (Exps. 1 and 3: 78% vs. 75%, 67%: $L = +2, -1, -1$, $F(1,68) = 0.43$, n.s. Exps. 2 and 4: 78% vs. 71%, 75%: $L = +2, -1, -1$, $F(1,68) = 0.23$, n.s.). These data indicate that SC algorithms detect cheaters just as well in novel social exchanges as they do in familiar ones. This is just what one would expect of a mechanism that builds frames.

The above social contract tests support the hypothesis that social contract algorithms guide inference for problems involving social exchange. The tests indicate that these algorithms are not only able to abstract social contract dimensions from unfamiliar situations, but they do so as efficiently in novel situations as in familiar ones.

Availability assessed: Does availability have any effect at all on familiar problems?

Although availability cannot explain the precisely patterned differences in performance among unfamiliar social contracts and all other problems, availability does appear to have had a minor, somewhat erratic effect on familiar descriptive problems.

In the literature, a standard test of the efficacy of availability is to compare a familiar descriptive problem to an abstract problem. The standard "now

you see it, now you don't" result of such experiments (see Cosmides, 1985, chap. 2) is mirrored in Experiments 1–4. The difference in percentage of falsifying responses between the familiar descriptive problem and the abstract problem is significant in Experiments 3 and 4 (Exp. 3: 50% vs. 12%: $F(1,23) = 13.80$, $p < .005$, $r = .61$. Exp. 4: 58% vs. 33%: $F(1,23) = 5.31$, $p < .05$, $r = .43$), but not in Experiments 1 and 2 (Exp. 1: 46% vs. 25%: $F(1,23) = 4.02$, n.s. Exp. 2: 38% vs. 29%: $F(1,23) = 1.31$, n.s.).

Using performance on the unfamiliar descriptive problem, instead of on the abstract problem, as a baseline for comparison, the familiar descriptive problem fares slightly better. The familiar descriptive problem elicited significantly more falsifying responses than the unfamiliar descriptive one in three out of the four experiments (Exp. 1: 46% vs. 21%: $F(1,23) = 5.31$, $p < .05$, $r = .43$. Exp. 2: 38% vs. 25%: $F(1,23) = 1.30$, n.s. Exp. 3: 50% vs. 12%: $F(1,23) = 13.80$, $p < .005$, $r = .61$. Exp. 4: 58% vs. 25%: $F(1,23) = 11.50$, $p < .005$, $r = .58$).

By considering the results of all four experiments ($n = 96$), we can estimate the relative sizes of the availability and social contract effects. Measured against the unfamiliar descriptive problems, the unfamiliar social contracts yield an effect size (r—see Rosenthal & Rosnow, 1984) of .70, whereas the familiar descriptive problems yield an effect size of .47. Overall, then, the social contract effect is 1.49 times the size of the availability effect—about half again as large.

These results can shed light on why performance with the transportation problem has been so erratic in the literature. When the transportation problem is measured against the AP—the standard test for availability in the literature—the effect size ($n = 96$) is only $r = .41$. Twenty-four (df 23) is a common sample size in the literature. Assuming the true effect size is .41, a sample size of 24 would yield an $F(1,23)$ of 4.65—barely over the $p < .05$ cut-off of 4.28. With just a little sample variation, one would sometimes get a significant effect, and sometimes not.

Summary, Experiments 1–4

Unfamiliar though they were, social contract problems reliably elicited social contract answers, even when these were radically at variance with formal logic. Furthermore, non-social contract problems (U-D, AP, F-D) did not show this distinctive pattern of variation. Availability alone can neither predict nor explain the results of these experiments. In addition to the social contract effect, there also appears to have been a marginal effect of availability on familiar descriptive problems.

Discussion for Part I

The social contract hypothesis systematically accounts for empirical results on the Wason selection task

The availability theories of reasoning maintain that the probability that one finds a content effect on the Wason selection task is directly proportional to the familiarity of the terms, rules and relations being reasoned about. In contrast, social contract theory maintains that humans have mental algorithms specialized for reasoning about social exchange, and that these algorithms determine how we reason on Wason selection tasks when their content involves social exchange. The specific features of the social contract algorithms were derived directly from modern evolutionary theory as it bears on social exchange, allowing a series of highly specific predictions to be made and tested.

Experiments 1–4 were designed to choose between two competing hypotheses:

(1) Availability is the sole determinant of performance on Wason selection tasks of varying content (the null hypothesis from the standpoint of most of the existing literature).
(2) Social contract algorithms are the major determinant of performance on Wason selection tasks when their content involves social exchange.

Six critical tests—comparisons for which social contract theory and availability theory make radically different predictions—were made by comparing performance on unfamiliar social contract problems with performance on both unfamiliar and familiar descriptive problems. Availability predicts a low percentage of logically falsifying, *P & not-Q*, responses for all unfamiliar rules, whether they are social contracts or not, and does not predict the response *not-P & Q* under any circumstance. In contrast, social contract theory predicts a high percentage of *P & not-Q* responses to "standard" social contracts, and a high percentage of *not-P & Q* responses to "switched" social contracts—no matter how unfamiliar the social contracts are. The critical tests were designed to unambiguously choose between social contract theory and the availability theories of reasoning. If social contract algorithms exist, then they should produce a highly distinctive and unusual pattern of results.

For all six tests, the social contract hypothesis was verified and the null hypothesis that availability is the sole determinant of responses was falsified. Each of these six tests was replicated, using different unfamiliar social contract problems. The six critical tests, and other experiments, established the following points:

(1) Unfamiliar *standard* social contracts elicit the predicted social contract response, *P & not-Q*, in the vast majority of subjects.
(2) Unfamiliar *switched* social contracts elicit the predicted social contract response, *not-P & Q*, in the vast majority of subjects.
(3) The percentage of social contract responses elicited by standard and switched social contracts is equivalent, even though these responses are quite distinct from a logical point of view (*P & not-Q* vs. *not-P & Q*). This is just what one would expect if the same algorithms were producing both responses.
(4) Social contract algorithms ignore for a switched social contract the cards they should choose for a standard one, and vice versa, just as social contract theory predicts.
(5) Social contract algorithms operate just as well in novel situations as they do in familiar ones: the percentage of social contract responses elicited by *unfamiliar* social contracts is equivalent to that elicited by *familiar* social contracts.
(6) Social contract algorithms are the major determinant of responses to problems whose content involves social exchange. More social contract responses are elicited by *unfamiliar* social contracts than falsifying responses by *familiar* descriptive problems. The social contract effect is almost 50% larger than the effect availability has on familiar descriptive problems.
(7) The social contract effect is replicable with a variety of familiar and unfamiliar social contracts.

The availability theories of reasoning presume the existence of innate learning mechanisms that are general-purpose and content-independent. However, no variant of availability theory can adequately explain the results of the experiments presented in this article. It is difficult to see how the association "cassava root–no tattoo" or "eats duiker meat–has never found ostrich eggshell" could have been "cued" from long-term memory (Griggs & Cox, 1982; Manktelow & Evans, 1979), let alone be the dominant association for over 70% of undergraduates tested (Pollard, 1982). No matter how wildly unfamiliar the rule's terms, social contract problems elicited social contract responses. Furthermore, if associations between specific terms were responsible for the pattern of results on social contract rules, then descriptive rules using the same unfamiliar terms should have elicited the same pattern; they did not. No theory whose predictive and explanatory power rests on associations between specific terms used in a social contract rule can explain the results of these experiments.

Availability theories that emphasize the role of mental modeling (Johnson-

Laird, 1982) or frames (Rumelhart & Norman, 1981; Wason, 1983) in recognizing *logical* contradiction cannot explain the following aspects of the results:

(1) Many more subjects "falsified" in response to the unfamiliar standard social contract problems than in response to the unfamiliar descriptive problems. Yet the scenarios described in each are culturally alien. Why would subjects find the scenarios described in the social contract problems so much easier to model than those described in the descriptive problems? Even worse, more people "falsified" in response to the *unfamiliar* social contracts than in response to the *familiar* descriptive problem. Why would an unfamiliar, culturally alien scenario be easier to model than a familiar one?

(2) Why would this situation reverse itself on switched social contracts, for which the scenario to be modeled is identical to that for the standard social contract? Unlike the standard social contracts, the switched ones do not elicit logically falsifying responses—although they do elicit the correct social contract response.

The only response an availability theorist of the modeling variety could make would be to agree with the analysis presented here that people do indeed have a *generalized* social contract "frame" that recognizes and operates on the cost–benefit structure of a social contract, and that has procedures that allow one to detect cheaters (but not logical contradiction), but to claim that this frame was acquired exclusively through "experience". More precisely, their claim must be that the social contract algorithms revealed in these experiments must themselves have been induced from *experience structured solely by the innate, content-independent, general-purpose information-processing systems presumed by associationists* (Fodor, 1983).

A claim very similar to this one was recently made by Cheng and Holyoak (1985) and Cheng, Holyoak, Nisbett, and Oliver (1986). Experiments testing their theory against social contract theory are presented in Part II of this article.

On implicit inference and deontic logic

For the social contract algorithms to operate, they must include the following two components:

(1) *An interpretive component.* Rarely, if ever, are all the contractual conditions that must be "mutually manifest" (Sperber & Wilson, 1986) to all exchange partners stated explicitly. Nevertheless, the interpretive component of the social contract algorithms must be able to recognize a

situation as one of social exchange, using a small but sufficient set of diagnostic cues. In this article, I have proposed that explicit information allowing the subject to perceive the rule as having the characteristic cost–benefit structure of a social contract constitutes a sufficient set of diagnostic cues. After having correctly categorized the situation as one of social exchange, the interpretive component must then map all explicitly described elements in the situation to their social exchange equivalents (cost–benefit representations, representations of the two agents, the obligation relationship, the entitlement relationship, and so on). To do this, implicit inference procedures must fill in all necessary steps—even those that have not been explicitly stated—fleshing out the semantic relations among the elements of the situation according to the principles of social exchange (Cosmides, 1985; Cosmides & Tooby, 1989).

(2) *A "look for cheaters" procedure.* The social contract algorithms must include a procedure that operates on the resulting representations to detect cheaters.

The experiments reported herein have primarily tested for the presence of the "look for cheaters" procedure; however, their design also involves, and to some extent tests, some features of the interpretive component's implicit inference procedures.

The most important example of this implicit inference is the subject's interpretation of the deontic relations in the rule itself. In a situation of social exchange, in order to be *entitled* to receive a benefit, one is *obligated* to provide a benefit, usually at some cost to oneself. These deontic concepts should be inferred by the interpretive component, *even when they are not explicitly mentioned in the rule.* The deontic operator "may", indicating entitlement, should be assigned to the benefit clause, and the deontic operator "must", indicating obligation, should be assigned to the cost clause, no matter what their position in the social contract rule. Thus, social contract theory predicts that when the interpretive component recognizes a term as a cost or benefit, it will cause the appropriate deontic operator to be assigned.

This means that cost–benefit rules that lack deontic operators, such as "If you take the benefit, then you pay the cost" and "If you pay the cost, then you take the benefit" can function as "projective" tests: because they mention no deontic operators, one can see which ones subjects spontaneously project onto the consequent clause of the rule (in English, deontic operators typically do not appear in the antecedent clause).

On this analysis, the interpretive component's implicit inference procedures should cause subjects to read a deontic "must" into the cost clause of

a standard social contract, even when it is not actually there ("If you take the benefit, then you (must) pay the cost"). This appears to be what subjects did: in Experiments 1 and 2, all four standard social contracts (two laws, two exchanges) elicited equally high percentages of *P & not-Q* responses, even though the deontic "must" was explicitly stated in only one of the four rules tested. This suggests that subjects were "reading in" a deontic "must" even when it was entirely absent. It also indicates that an explicitly stated "must" is not, by itself, a diagnostic cue of social exchange: it is not a necessary cue, and, as the experiments in Part II will show, it is not a sufficient cue—permission rules that have a "must", but that lack the cost–benefit structure of a social contract, do not elicit the same results.

Similarly, the interpretive component should cause subjects to read a deontic "may" into the benefit clause of a switched social contract, even when it is absent ("If you pay the cost, then you (may) take the benefit"). None of the four switched social contracts tested in Experiments 3 and 4 had an explicit "may" in the benefit clause, yet the percentage of correct social contract responses (*not-P & Q*) was just as high as in Experiments 1 and 2. This indicates that an explicitly stated "may" is not necessary for one to recognize a situation as one of social exchange: the interpretive component can recognize a switched social contract as such, even when it lacks this deontic operator.

Because one may take a benefit one is entitled to, but one is not required to do so, reading a "must", rather than a "may", into the benefit clause of a switched social contract would violate the principles of social exchange. This would yield "If you pay the cost then you (must) take the benefit", which would require a *P & not-Q* response. Even though it would usually be natural to read a modal "must" into the consequent clause of a conditional rule, it is clear that subjects did not do so for the switched social contracts: only 1 out of the 48 subjects in Experiments 3 and 4 answered *P & not-Q*. This suggests that the interpretive component's implicit inference procedures were blocking the usual insertion of a "must", which is just what would happen were they substituting a "may".

Indeed, if subjects were inserting implicit "mays" and "musts" according to some non-social contract pattern (e.g., "If you must take the benefit then you may pay the cost"), or else not inserting them at all, then their responses would not have been consistent with the social contract predictions.

I have not dwelled on the operation of such predicted implicit inferences in these experiments, because their presence is not by itself persuasive. This is because, unlike the structure of the "look for cheaters" procedure, there is no normative theory of interpretation that is sufficiently well specified to test against social contract theory. I believe all ordinary readers find the

implicit insertions of the "musts" and "mays" in the appropriate (i.e., "social-contract-appropriate") locations of the rules entirely natural, even inevitable, and I would claim that this "naturalness" is due to the operation of the social contract algorithms. However, although this is a clear prediction of social contract theory, in the absence of other theories of interpretation with strong contrary predictions, it is not clear that another theory would lead one to expect something else. Even if skeptics cannot state in a predictive manner exactly why they feel such an interpretation is natural, the fact that it does seem natural and uncontroversial makes its prediction interesting, but inadequate by itself, as a critical test of social contract theory.

For example, Manktelow and Over (1987), in their stimulating review of human reasoning, ask whether people have a mental deontic logic: mental rules of inference governing moral obligation and entitlement. Of course, social contract theory proposes that, for contexts involving social exchange, we do have mental rules of inference governing moral obligation and entitlement (see "Social contracts as speech acts" in Cosmides, 1985, or Cosmides & Tooby, 1989). These could be thought of as embodying a (highly circumscribed) deontic logic. Given these parallels, does deontic logic constitute an alternative explanation for the data?

Given the current state of development of deontic logic, the answer would seem to be no. Deontic logic and social contract theory appear to differ in their domain of operation, and perhaps even in their rules of inference. Deontic logic is rooted in the concepts of permission and obligation, and should operate wherever those concepts are found. Social contract theory is far more limited in its scope, and instead predicts that the "look for cheaters" procedure will operate only when a rule has the cost–benefit structure of a social contract. This implies that a rule that lacks this cost–benefit structure, but that otherwise implies permission or obligation, will not elicit the effect. As we will see in the experiments of Part II, this is exactly what happens. Thus, the very pattern of results in Part I that suggests the presence of a deontic logic does not obtain for all deontic rules, but only for those having the cost–benefit structure of a social contract. However, Manktelow and Over argue that deontic logic cannot yet be explicitly tested as a counter-hypothesis to social contract theory, because philosophers are still arguing about what it would predict. For example, they point out that it is not clear whether, for switched social contracts, deontic logic predicts the *not-P & Q* response obtained, or whether it predicts that the subject should turn over no cards whatsoever. Furthermore, to be made into a psychological hypothesis, the rules of a formal deontic logic would have to be combined with an interpretive schema that could identify when they should be invoked. A simple, syntactic interpretive schema that activates deontic logic in re-

sponse to explicit deontic operators cannot explain the data herein, as the social contract effect obtained even when the rules tested had no explicit deontic operators.

A less empirical reason for preferring social contract theory to deontic logic as an explanation for these results is that there is an explanation of why the mind should contain social contract algorithms, while there is no explanation of why it should contain a deontic logic. Its presence would be particularly puzzling given the apparent absence of the propositional calculus: why have access to a formal logic of limited applicability, but not to a formal logic which can operate at the same level of abstraction to produce generally useful knowledge?

Social contract theory is directly derived from what is known about the evolutionary biology of cooperation, and is tightly constrained as a result. It explains why it should be present in the human mind, what its domain of operation will be, what kinds of implicit inferences it will generate, and what the structure of the "look for cheaters" procedure will be. Given that social contract theory and deontic logic both deal with concepts of entitlement and obligation, there are clearly recognizable similarities between the two, and their similarities or differences will become clearer once deontic logic reaches the level of specificity that social contract theory has. One certainly cannot deny the possibility that in the future someone may produce a theory that accounts for these same data, by supplying the missing interpretive component and by deriving a better-specified deontic logic than exists today. However, it seems more likely that the explanation for their convergence is that philosophers construct deontic logics by drawing on intuitions generated by their adaptive logics, including the logic of social exchange. This would explain not only their resemblance, but also the source of the difficulties encountered in their formalization. Such logics develop "inconsistencies" (i.e., seem unreasonable) when philosophers attempt to generalize them beyond the contexts in which they evolved to operate.

PART II: PERMISSION SCHEMAS OR SOCIAL CONTRACT ALGORITHMS? CHENG AND HOLYOAK'S "PRAGMATIC REASONING SCHEMAS"

Cheng and Holyoak (1985) and Cheng et al. (1986) propose that humans reason about "realistic" situations using sets of generalized, relatively abstract production rules which they call "pragmatic reasoning schemas". The name is well chosen: these knowledge structures are "schemas" because they are schematic—generalized and relatively abstract; they are "reasoning" schemas

because they consist of production rules, and they are "pragmatic" because they are activated by domains that are defined functionally, by classes of goals. Cheng and her colleagues believe that these schemas are produced by induction, as it operates over goal-defined domains.

For example, Cheng and Holyoak (1985) propose that people have a "permission schema" for reasoning about "regulations ... imposed typically by an authority to achieve some social purpose" (p. 398). The permission schema is hypothesized to consist of the following four production rules:

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

Cheng and Holyoak maintain that most of the thematic problems that have elicited high levels of logical falsification on the Wason selection task can be characterized as permission rules, expressed in the linguistic format of Rule 1. I would have to agree with them, insofar as all social contracts fit the permission schema (however, not all permission rules are social contracts, as will be discussed below).

When the above four production rules operate on a permission rule, they produce, by coincidence, the logically falsifying response, *P & not-Q*. Rule 1 causes one to choose the *P* card, and Rule 4 causes one to choose the *not-Q* card. Rules 2 and 3 cause no card to be chosen: because their consequents admit of possibility, rather than certainty, any result is consistent with them, and, therefore, no card can falsify them. Consequently, permission rules elicit a "content effect" on the Wason selection task: people appear to reason logically when the task's content involves "permission". Cheng et al. (1986) make a similar argument about a similar set of production rules, which they call an "obligation schema". An obligation schema employs rules embodying the same four modal relations; however, the representations differ slightly, as in "Rule 1: If condition C occurs, then action A must be taken". The main difference between the two schemas is a time relation. If the action that the rule obligates one to take must be done first (i.e., if it is a *pre*condition), then the permission schema is activated; if the rule allows that action to be done second, then the obligation schema is appropriate. (Because this difference is so minor, I will use "permission schema" to mean either permission schema or obligation schema, unless otherwise specified.) Cheng and her colleagues also propose the existence of covariation schemas and causal schemas.

Social contract theory and pragmatic reasoning theory agree on several points. First, both theories agree that people lack a "mental logic" (Johnson-

Laird, 1982). In other words, they agree that the innate architecture of the human mind does not include a set of algorithms that instantiate the rules of inference of the propositional calculus. Second, both maintain that in solving the selection task, people use rules of inference appropriate to the domain suggested by the problem, and that these rules of inference may be different for different content domains. Third, both theories propose algorithms specialized for reasoning about social exchange—as we shall see, all social contracts are either permission or obligation rules, but not all permission and obligation rules are social contracts.

Despite these similarities, social contract theory and pragmatic reasoning theory differ in two important respects: (1) the structure of the proposed algorithms, and (2) their origin. Cheng et al. believe that their schemas originate in induction, that their rules of inference are the product of "experience" structured only by innate information-processing mechanisms that are domain-general. In contrast, social contract theory proposes that the social contract algorithms' rules of inference are themselves innate, or else the product of "experience" structured by innate algorithms that are domain-specific.

As it stands, Cheng et al.'s claim regarding the origin of their schemas is unfalsifiable. Therefore, the two theories' differing contentions regarding the origin of the proposed algorithms is not directly testable. However, these theories are subject to plausibility arguments based on existing data, which will be made in the discussion for Part II.

Fortunately, however, the proposed differences in the structure of social contract algorithms and permission schemas are testable. Experiments testing the most important difference in structure—the proposed level of representation and domain of operation—are presented herein.

Different levels of representation: Actions and preconditions or costs and benefits?

A social contract specifies what two or more individuals intend to exchange. In a social contract, in order to be entitled to receive a benefit from a person, you are required to provide that person with something that he or she considers to be a benefit; your providing that benefit usually (but not necessarily[8])

[8]The necessary and sufficient cost-benefit restrictions on a social contract are spelled out in detail in Cosmides (1985) and in Cosmides and Tooby (1989). Here, suffice it to say that in social exchange it is not strictly necessary that each side suffer a cost in the course of providing a benefit to the other side (although this will usually be the case); what is essential is that each side be provided with a benefit. This providing of a benefit to the other party is *required*, and usually (although not necessarily) entails a cost; hence in this paper the term specifying this provision is referred to as the cost or requirement term. Situations in which provision of a benefit does not entail a cost are discussed in the sections of Cosmides and Tooby (1989) on "baseline fraud".

entails your incurring a cost. Thus, a social contract rule relates *perceived benefits* to *perceived costs*, expressing an exchange in which an individual is obligated to pay a cost to an individual or group in order to be entitled to receive a benefit from that individual or group.

By hypothesis, social contract algorithms represent the world at a different level of abstraction than do permission and obligation schemas. Social contract algorithms represent the world in terms of "benefits" and "costs": increases and decreases in utility for the actors involved in the exchange (see Table 5). These benefits and costs are defined with respect to an individual's "zero level utility baseline": that individual's utility (including expectations about the future) at the time the offer is made, but independent of it. Certain social laws, the simultaneous trading of goods and services, and "reciprocal altruism"—the informal, non-simultaneous exchange of favors—are all forms of social exchange, and can be expressed in terms of this cost–benefit structure. In contrast, permission and obligation schemas represent the world in terms of "actions to be taken" and "(pre)conditions to be satisfied" ("preconditions" for permission schemas, "conditions" for obligation schemas).

All social contract rules involve permission (or, more strictly, entitlement), but not all permission rules are social contract rules. This is because the statement "If one is to take the benefit, then one must pay the cost"—a social

Table 5. *Sincere social contracts: cost–benefit relations when one party is sincere, and that party believes the other party is also sincere*[a]

	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	Positive:	Positive:	Positive:	Positive:
margin	B(me) > C(me)	B(you) > C(you)	B(me) > C(me)	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)"	

[a]B(X) = benefit to X; C(X) = cost to X; 0(X) = no change in X's zero level utility baseline. The zero-level utility baseline is the individual's level of well-being (including expectations about the future) at the time the offer is made, but independent of it. Benefits and costs are increases and decreases in one's utility, relative to one's zero-level utility baseline.

contract—is subsumed by the statement "If one is to take action A, then one must satisfy precondition P"—a permission. However, the reverse is not true. All "benefits taken" are "actions taken", but not all "actions taken" are "benefits taken". *A permission rule is also a social contract rule only when subjects interpret the "action to be taken" as a rationed benefit, and the "precondition to be satisfied" as a cost requirement.*

This hypothesized difference in level of representation has empirical consequences. According to permission schema theory, permission rules that are *not* social contracts should elicit a content effect on the Wason selection task; according to social contract theory, they should not. By hypothesis, then, permission schemas operate over a larger domain than do social contract algorithms. Social contract algorithms should be activated only by social contract rules. Permission schemas should be activated not only for social contract rules, but also for non-social contract (non-SC) permission rules. From the point of view of social contract theory, action–precondition representations are over-general.

Therefore, it is possible to construct critical tests that allow one to decide which kind of representation is psychologically real: the action–precondition representation, or the benefit–cost representation. Unfortunately, Cheng et al.'s experiments do not allow one to decide the issue because all the permission rules they tested were, by coincidence, social contracts.

As discussed in Part I, Cheng and Holyoak's first experiment tested two rules: "If a letter is sealed, then it must carry a 20-cent stamp", and "If the form says 'ENTERING' on one side, then the other side includes cholera among the list of diseases". Each subject was tested on both rules; however, one of the rules had been given a "social purpose" through the addition of contextual information, whereas the other rule had been stripped of this context.

Although this was no part of their theory, it so happens that the "social purposes" that Cheng and Holyoak chose conferred a clear cost–benefit structure on the rules, thereby making them into social contracts. For example, the text surrounding the postal rule explained:

> The rationale for this regulation is to increase profit from personal mail, which is nearly always sealed. Sealed letters are defined as personal and must therefore carry more postage than unsealed letters.

In other words, noting that people prefer to seal their personal letters, the post office decides to ration this benefit by charging more for it, thereby benefiting itself. In cost–benefit terms, the postal rule therefore translates to, "If you want the benefit of sealing a letter, then you must pay a cost for it". The context for the other rule, which was set in an immigration office, stated

that the form listed inoculations that the passenger had had in the past 6 months, and that the purpose of the rule was "to ensure that entering passengers are protected against the disease." The statement about inoculations labels the consequent as a cost, as most people know that inoculations are painful and frequently cause several days of illness. Furthermore, most people know that countries consider entry within their borders a privilege that they have the right to regulate. Thus, the context provided for this rule makes it clear that the passenger wishes to enter the country, that this is a privilege that the country rations, and that the price one must pay for entrance is a painful inoculation, or, in cost–benefit terms, "If you want the benefit of entering our country, then you must pay the price of being inoculated." Eighty-five to 90% of subjects chose *P & not-Q* for the rules with a clear, contextually defined cost–benefit structure, compared to 55–60% of subjects for the no-context rules.

Even the no-context rules were close analogs to real-life social contract rules, which could explain why performance on them was as high as 60%. Although American subjects are used to thinking of sealing an envelope as a free good, rather than a rationed benefit, they do know that postage is a cost of mailing, and that the post office treats other properties of letters, such as their weight, class and destination, as rationed benefits that require more postage. Similarly, even without the added information about inoculations, most people know that diseases are of interest to immigration officers because entry is a privilege that can be denied to people who lack the appropriate inoculations. From this general world knowledge, it is a small leap to the social contract interpretation. Moreover, this leap would have been facilitated by exposure to the other rule, which had a clear cost–benefit structure, and therefore would have activated the subjects' social contract algorithms: remember, Cheng and Holyoak allowed their subjects to flip back and forth between the two problems, changing answers. As will be discussed below, experiments by Cox and Griggs (1982) indicate that previous exposure to a social contract rule can facilitate performance on a rule that has constituents suggestive of a social contract. This was probably a significant factor boosting performance on Cheng and Holyoak's no-context rules, because in an experiment which lacked this confound, Griggs and Cox (1982) found that the same no-context postal problem elicited a *P & not-Q* response from only 4% of their subjects. Given their choice of rules and this methodological confound, Cheng and Holyoak themselves concluded that their subjects were "sometimes able to provide their own rationales" for the no-context rules. Hence, a social contract theorist could characterize this experiment as one in which a clear social contract was compared to a rule that, with very little prompting, could also be interpreted as a social contract, and note that the

clear social contract elicited more social contract responses than the ambiguous one.

The content-free permission rule tested in Cheng and Holyoak's second experiment was also a social contract ("If one is to take action 'A', then one must first satisfy precondition 'P'. In other words, in order to be permitted to do 'A', one must first have fulfilled prerequisite 'P' "). Although this content-free permission problem does not mention costs and benefits by name, it still has an implicit cost–benefit structure. After all, saying that one must fulfill or satisfy a precondition in order to be permitted to do something is just another way of saying that one must pay a cost or meet a requirement. Furthermore, people do not tend to insist that such conditions be fulfilled *before* an action is taken unless they fear that they will not be fulfilled *after* it is taken; one would not guard against such cheating unless the condition were something the person would rather not do—i.e., a cost. In addition, saying that someone is *permitted* to take action A linguistically marks "action A" as a rationed benefit: it implies that the person *wants* to take action A (your mother permits you to get ice cream, she does not "permit" you to be spanked), and it implies that the person doing the permitting has the power to forbid action A. The inference that "action A" is a benefit stands even without the paraphrase about permitting: in real life, people don't bother to make rules preventing one from taking an action that one does not want to take. The only way to "make sense" of a rule such as "If you want to eat dirt, then you must give me $10", is to assume that the person to whom it is addressed *wants*, for some inexplicable reason, to eat dirt (in fact, Fillenbaum (1976) found that people consider rules of this kind "extraordinary"). To avoid a social contract interpretation, the context of the rule would have to lead the subject to assume that one is no better off by taking action A than by not taking it, and no worse off by satisfying precondition P than by not satisfying it. By including a paraphrase about "permitting", this experiment did quite the contrary.

Thus, Cheng and Holyoak's experiments do not test a permission rule that has a cost–benefit structure against one that does not. Therefore, these experiments do not allow one to decide whether the effect was produced by social contract algorithms or by a permission schema.

A number of permission rules that lack a cost–benefit structure have been tested in the Wason selection task literature, and these have not elicited content effects. For example, Griggs and Cox (1983) systematically negated components of two standard social contracts: D'Andrade's Sears problem and their own drinking-age problem. Some of these negations preserved the cost–benefit structure of a standard social contract, whereas other, similarly negated rules, did not. These latter rules—call them "deformed" social contracts—had components that are recognizable as costs and rationed benefits.

However, these components were arranged such that they *violate* the principles of social exchange, wherein one is obliged to pay a cost in order to be entitled to a benefit. For example, "If you take the benefit, then you must *not* pay the cost" is a deformed social contract. Because they violate the principles of social exchange, which, by hypothesis, are instantiated in the social contract algorithms, subjects should feel that deformed social contracts do not "make sense".

Both deformed and standard social contract rules are permission rules: both relate "actions" to "preconditions" in a permission format. Therefore, according to Cheng and Holyoak's theory, subjects should choose the "action to be taken" card (*P*) and the "precondition not met" card (*not-Q*) for both sets of rules, regardless of their cost–benefit structure. In contrast, social contract theory predicts that subjects will choose these cards for standard social contracts, but not for deformed social contracts. This is because these cards represent a "benefit accepted" (*P*) and a "cost not paid" (*not-Q*) for standard social contracts, but not for deformed ones.

The results support social contract theory. Subjects chose the "action to be taken" card (*P*) and the "precondition not met" card (*not-Q*) for standard social contracts, but not for deformed ones, even though both were permission rules (for detailed discussion, see Cosmides, 1985). This was true even when the standard social contracts had very complicated and abstruse structures of negation. Permission schema theory cannot readily account for this result. As long as a permission rule relates actions and preconditions, subjects should choose the "action to be taken" card and the "precondition not met" card. The cost–benefit status of these cards should not affect performance.

Another permission rule that lacked a clear cost–benefit structure was tested by Cox and Griggs (1982). They tested an "apparel-color" problem—"If a person is wearing blue, then that person must be over 20 years old"—to see if people could reason by analogy to the drinking-age problem, which is a true social contract ("If a person is drinking beer, then that person must be over 20 years old"). The apparel-color rule is not a social contract because in our culture, "wearing blue", as opposed to "wearing green", is not a rationed benefit. The apparel-color rule is, however, a permission rule: it fits the "If one takes action A, then one must satisfy precondition P" representation of a permission. Yet when it was administered *before* the drinking-age problem, only 25% of subjects falsified. It elicited a substantial content effect only when administered *after* the drinking-age problem, which would have suggested a cost–benefit interpretation for the apparel-color rule.

Although this result is difficult for permission schema theory to explain, it is consistent with social contract theory. Cox and Griggs attributed the transfer effect to "reasoning by analogy". Yet solving the drinking-age problem

did not enhance performance on an abstract problem. To be predictive, explanations that invoke reasoning by analogy require *theories of analogy*: theories specifying which dimensions subjects use for construing similarity. A Darwinian algorithms perspective can provide the theories of analogy for various domains that would make "reasoning by analogy" predictive. For the domain of social exchange, for example, the easier it is to interpret a rule as having the cost–benefit structure of a social contract, the more likely it is to elicit a social contract response. Ease of interpretation should be a function of: (1) whether the social contract algorithms are activated; (2) how similar the constituents are to costs and benefits that the subject is familiar with; and (3) whether this interpretation would yield a well-formed social contract. This would explain why the apparel-color problem elicited few social contract responses when it came before the drinking-age problem, but many when it came after. The drinking-age problem has a clear cost–benefit structure, which would have activated the subjects' social contract algorithms. The two rules share the same consequent ("being over 20"), thus defining the apparel-color problem's consequent as a cost requirement. And drinking beer is well known as an age-related benefit. Given that the two rules share the same consequent, it would therefore be reasonable to assume that "wearing blue" might also be an age-related benefit in some unusual circumstance. In a similar vein, Cheng et al. (1986) have shown that training subjects in the relations involved in obligation produces a content effect in problems that can be interpreted as social contracts, whereas training in logical relations does not.

Other non-social contract permission rules have been tested also. As mentioned earlier, when subjects view sealing an envelope as a free good (and when the experimental design prevents reasoning by analogy to a social contract problem), the post office problem described above does not elicit a content effect, even though it is a permission rule (Golding, 1981; Griggs & Cox, 1982). Neither do permission problems relating letters and numbers, such as those tested in Part I of this article (the prescriptive AP problems), as well as those tested by D'Andrade in Rumelhart and Norman (1981) and Cosmides (1985, Exp. 5).

Although these results do not help permission schema theory, they do not falsify it either. Cheng and Holyoak could argue that these permission rules did not produce a content effect because they failed to activate the permission schema. Citing their first experiment, in which the rule with a social purpose elicited a larger content effect than the rule without one, they could claim that a rule must have a "social purpose" if it is to activate the permission schema.

In pragmatic reasoning theory, "purposes" serve to discriminate schemas

at the interpretive stage. They determine which schema—permission, obligation, covariation, causal, etc.—will be activated by a problem. Cheng and Holyoak do not believe that an action–precondition representation is always sufficient to activate the permission schema; sometimes a social purpose is necessary (they never say whether the representation alone is sometimes sufficient, and, if so, when). In contrast, social contract theory holds that recognizing a rule as having the cost–benefit representation of a social contract is sufficient to activate the social contract algorithms; the rule need not have a "social purpose".

In order to be predictive, permission schema theory should define what counts as a social purpose, and provide criteria that allow one to discriminate one kind of purpose from another. Will any cover story that makes the rule sound plausible do? (If so, what are the criteria that determine "plausibility"?) Must the purpose specify how society, or the institution making the rule, benefits by having people follow the rule? Alternatively, must the story contain elements that make the rule's terms interpretable as costs and benefits in the format of a social contract, as social contract theory maintains? Unfortunately, Cheng and Holyoak never specify what counts as a "social purpose", forcing one to rely on hazy intuitions.

Yet without such theoretical specification, one does not know whether or not a problem's context has given a rule a social purpose, and one cannot predict which schema will be activated by which purpose. For example, in Experiment 1 (Part I), the social contract law regarding duiker meat states a rule, but it never says why the elders would wish to make such a rule in the first place. Did the context provide a social purpose? If not, then permission schema theory cannot explain why this rule elicited such a large content effect. If it did, then what was the social purpose?

Even more problematic are the private exchanges tested in Experiments 2 and 4. These are private arrangements, not "regulations imposed by an authority"—Cheng and Holyoak's gloss of a permission. And although these exchanges were to benefit the individuals who agreed to them, they were not designed to benefit "society" or any institution: they were not "imposed to achieve a social purpose". Nevertheless, they elicited a large percentage of social contract responses.

The provision of a social purpose is a pivotal aspect of Cheng and Holyoak's theory. Without it, they cannot explain why the non-SC permission rules already tested in the literature failed to produce a content effect. But if they insist that a social purpose is necessary, then they cannot explain why the social contracts tested in Part I elicited large content effects: the private exchanges (and, arguably, the social contract laws) had no "social purpose".

In sum, a number of experiments suggest that the action–precondition

representations of permission schema theory are over-general, and that, as social contract theory predicts, content effects are elicited only by rules having the cost–benefit structure of a social contract. Yet none of these experiments provide a clean critical test to differentiate permission schema theory from social contract theory.

The central issue is: which kind of representation is psychologically real—the action–precondition representation of permission schema theory or the benefit–cost representation of social contract theory? To decide this issue, one would have to test a permission rule that lacks the cost–benefit structure of a social contract, but has a social purpose. Performance on this rule must be compared to performance on a rule that has the cost–benefit structure of a social contract. This is exactly the test that was performed in Experiments 5 and 6 below. If social contract theory is correct, then the social contract rule will elicit many more social contract responses than will the non-social contract permission rule. If permission schema theory is correct, both rules will elicit a high percentage of social contract responses.

Experiments 5 and 6

Social contract theory and pragmatic reasoning theory differ in their hypotheses about the representational structure of the schemas people use to reason about social contracts and/or permissions. Because of this, they also differ in the scope of the domains over which they operate. The purpose of Experiments 5 and 6 was to test a standard social contract against a non-social contract (non-SC) permission rule that has a social purpose. A non-SC permission rule is a rule that lacks the cost–benefit structure of a social contract, but that does fit the action–precondition representation of a permission. If social contract theory is correct, then the standard social contract will elicit a high percentage of the predicted social contract response, *P & not-Q*, but the non-SC permission rule will not. If permission schema theory is correct, then the percentage of *P & not-Q* responses elicited by the non-SC permission rule will be as high as that for the social contract rule.

Subjects

Eighty students from Stanford University participated in Experiments 5 and 6, 40 in each experiment. All subjects were paid volunteers recruited by advertisement. Experiment 5 had 12 females and 28 males (mean age: 18.5 years (no data on age of 2 subjects)); Experiment 6 had 17 females and 22 males (no data on sex of 1 subject; mean age: 19.9 years (no data on age of 3 subjects)).

Materials and procedures

The procedure for Experiments 5 and 6 was identical to that for the experiments of Part I, except the experimenter did not read the instructions aloud to subjects. Each test booklet had two Wason selection tasks: a standard social contract and a non-SC permission rule. For half the subjects, the social contract problem came first; for the other half, the non-SC permission rule came first. Only the first problems allow a clean critical test of the two hypotheses; the second problem was included for a future study on transfer effects. All experiments in Part II had a between-subjects design.

Materials for Experiment 5: The rules used for both the social contract problem and the non-SC permission problem were: "If a student is to be assigned to Grover High School, then that student must live in Grover City", and "If a student is to be assigned to Milton High School, then that student must live in the town of Milton." Each booklet contained either the "Grover High" version of the social contract problem and the "Milton High" version of the non-SC permission problem, or vice versa; this variable was counterbalanced across subjects.

The social contract problem and non-SC permission problem differed only in surrounding story context; the rules used were identical. The surrounding story for the social contract problem explained that being assigned to Grover High is a benefit (compared to being assigned to Hanover High), and that living in Grover City is a cost (compared to living in Hanover). The surrounding story for the non-SC permission problem gave the rule a social purpose (following the rule will allow the Board of Education to develop the statistics it needs in order to assign the proper number of teachers to each high school), but it did not define the terms of the rule as benefits or costs (value differences between schools or places are irrelevant to the stated purpose, and none are mentioned; being assigned to Grover or Hanover High are thereby portrayed as of equal value, as is living in Grover City vs. Hanover). Each problem contained several suggestions that the rule may not have been followed. In the case of the non-SC permission problems, it was suggested that the person responsible for following the rule may have violated it by mistake (because the rule is not a social contract, the person following it has nothing to gain by cheating), whereas in the social contract problems it was suggested that the person may have violated it through intentional cheating. From the point of view of permission schema theory, subjects should be able to detect violations of a permission rule regardless of whether they were caused by carelessness or by cheating. The full text of all problems used is reported in the Appendix.

Experiment 6: The procedure for Experiment 6 was identical to that de-

scribed for Experiment 5. The materials, however, were different. Instead of having a culturally familiar setting (rules in a local Board of Education), the rules were set in the same fictitious cultures described in Experiment 1. The purpose of this experiment was to replicate the results of Experiment 5 in a culturally alien setting.

Materials for Experiment 6: The rules used for both the social contract problem and the non-SC permission problem were: "If a man eats cassava root, then he must have a tattoo on his face", and "If you eat duiker meat, then you have found an ostrich eggshell". The social contract problems were the two unfamiliar standard social contracts used in Experiment 1. The non-SC permission problems gave these rules a social purpose, but did not define their terms as costs and benefits. In both cases, the social purpose was one that would benefit the whole group: the purpose of the rule was to ration two staple foods that the tribe depended on, so that the foods would not become extinct through overuse, and the tribe with them. However, an individual was not *differentially benefited* by eating one of the staple foods versus the other. For example, in the cassava root problem, the two staple foods were of equal value (e.g., molo nuts were just as tasty and nutritious as cassava root), and having a tattoo was not worse than not having one (it simply marks what clan one is in). Therefore, eating cassava root was not a benefit (compared to eating molo nuts), and having a tattoo was not a cost. The same was true of the duiker meat problem. The full problem texts are reported in the Appendix.

Results

Percentage *P & not-Q* responses:

	Social contract prediction: Social contract > non-SC permission		Permission schema prediction: Social contract = non-SC permission	
	high	low	high	high
Exp. 5:	75%	> 30%		
Exp. 6:	80%	> 45%		

Just as social contract theory predicts, the social contract problems elicited far more *P & not-Q* responses than the non-SC permission rules, which lacked a cost–benefit structure (Exp. 5: 75% vs. 30%: $Z = 2.85$, $p < .0025$, $\phi = .45$. Exp. 6: 80% vs. 45%: $Z = 2.29$, $p < .0115$, $\phi = .36$). This was true even though the non-SC permission rules had a "social purpose". Permission schema theory does not predict, and cannot account for, this result.

Experiment 7

Does having a social purpose enhance performance *at all* for a non-SC permission problem? To see, I administered the school problem to 20 more undergraduates, but the social purpose was left out of this version (see Appendix). (To remove the social purpose from the fictitious culture problem would not have been a fair test; it would have left the problem totally devoid of sense.)

According to social contract theory, the only reason that the social purposes in Cheng and Holyoak's first experiment enhanced performance was because they conferred the cost–benefit structure of a social contract onto the permission rules tested. If this interpretation is correct, then there should be no difference in performance between the no-purpose school problem and the school problem with a social purpose, because the social purpose used did not confer a cost–benefit structure on the school rule. However, if permission schema theory is correct, then more *P & not-Q* responses should be elicited by the problem that has a social purpose than by the problem that lacks one.

Subjects

Twenty students from Stanford University participated in Experiment 7, 10 females and 10 males (mean age: 18.9 years, no data on age of 1 subject).

Materials and procedures

The procedure was identical to that of Experiment 5. Each subject solved two problems. The first was a school problem from which the social purpose had been removed, and the second was a school problem with the social purpose intact.

Results

There was no significant difference in percentage of *P & not-Q* responses between the problem that lacked a social purpose and the problem that had one: 20% for the no-purpose problem versus 15% for the one with a social purpose ($F(1,19) = 1.0$, repeated measures). This is true even if one controls for problem position: the social purpose school problem in Experiment 5 was administered first, just as the no-purpose problem of Experiment 7 was, yet having a social purpose made no significant difference in performance (no-purpose: 20%; social purpose: 30%, $Z = 0.73$). Thus, a social purpose that did not confer the cost–benefit structure of a social contract onto the permission rule failed to enhance performance.

Experiments 8 and 9

In Part I, Experiments 3 and 4 showed that switched social contracts—ones of the form "If you pay the cost, then you take the benefit"—elicited the predicted social contract answer, *not-P & Q*. But what response does permission schema theory predict for switched permission rules?

Cheng and Holyoak only tested permission rules in the format of Rule 1 of the permission schema: "If the action is to be taken, then the precondition must be satisfied". However, the other three rules of the permission schema should be able to activate the schema as well. The switched form of Rule 1 is Rule 3: "If the precondition is satisfied, then the action may be taken". As before, Rule 1 would cause the "action is taken" card to be chosen, and Rule 4 would cause the "precondition has not been satisfied" card to be chosen (again, Rules 2 and 3 cause no card to be chosen). However, for a permission rule in the format of Rule 3, the "action is taken" card corresponds to the logical category *Q*, and the "precondition not met" card corresponds to the logical category *not-P*. Therefore, a switched permission rule should yield the same response as a switched social contract: *not-P & Q*. In Experiments 8 and 9, performance on switched social contracts was compared to performance on switched non-SC permission rules that had a social purpose.

In essence, switched social contracts have a "Rule 3" format; the only difference is that the "may" was left implicit in Experiments 3 and 4 of Part I. Because social contract theory hypothesizes that the social contract algorithms instantiate the principles of social exchange, subjects should be able to "read in" the correct deontic operator, even when it is omitted (see "On implicit inference and deontic logic", in the Discussion for Part I).

Implicit "mays" should present no problem for permission schema theory: using permission rules that happened to be social contracts, Cheng and Holyoak themselves showed that subjects frequently paraphrase rules of the form "The benefit is to be taken only if the cost is paid", as "If the cost is paid, then the benefit *can* be taken" (Cheng & Holyoak, 1985: Experiment 3).

As long as it grants the implicit "may", permission schema theory makes the same prediction for switched social contracts as social contract theory does. This is because "If you pay the cost, then you (may) take the benefit" is subsumed by "If the precondition is satisfied, then the action may be taken", Rule 3. (If permission schema theory does not grant the implicit "may", then the results of Experiments 3 and 4 already falsify it. Without the implicit "may", the switched social contracts tested have the format of Rule 1, for which the predicted response is *P & not-Q*. Yet only 1 out of 48 subjects made this response in Experiments 3 and 4: most chose *not-P & Q* instead.)

Thus, to distinguish permission schema theory from social contract theory, switched social contracts must be tested against switched permission rules that lack the cost–benefit structure of a social contract. This was done in Experiments 8 and 9. Social contract theory predicts that only the switched social contract rules will elicit a high percentage of *not-P & Q* responses; permission schema theory predicts this response from both switched social contracts and switched non-SC permission rules. All rules had the general form: "If the precondition is satisfied, then the action is to be taken". However, for the social contract rule, the precondition was a cost and the action a benefit, whereas for the non-SC permission rule they were not.

The "may" was left implicit for all rules. This is because "you *may* take action A" is just another way of saying "you are permitted to take action A": it linguistically marks "action A" as a benefit. Including the "may" would therefore defeat the purpose of the experiment, which was to test a switched permission rule that *lacks* a cost–benefit structure. The omission of the "may" should not be a problem for pragmatic reasoning theory, however, since the point of Cheng and Holyoak's third experiment was to show that subjects spontaneously supply a modal "can" when paraphrasing permission rules into their switched form. After all, if subjects can read a deontic "may" into a switched social contract (which permission schema theory must admit in order to explain the results of Part I), they ought to be able to read it into other switched permission rules as well.

Experiment 8 tested a switched version of the school problem, and Experiment 9 tested a switched version of the fictitious cultures problem.

Subjects

Eighty students from Stanford University participated in Experiments 8 and 9, 40 in each experiment. All subjects were paid volunteers recruited by advertisement. Experiment 8 had 13 females and 27 males (mean age: 19.1 years); Experiment 9 had 16 females and 24 males (mean age: 18.7 years (no data on age of 2 subjects)).

Materials and procedures

The procedure was identical to that of Experiments 5 and 6. The materials were also identical, except that the rules were "switched". The school rules tested in Experiment 8 were, "If a student lives in Grover City, then that student is to be assigned to Grover High School", and "If a student lives in the town of Milton, then that student is to be assigned to Milton High School". The fictitious culture rules tested in Experiment 9 were, "If a man

has a tattoo on his face, then he eats cassava root", and "If you have found an ostrich eggshell, then you eat duiker meat".

Results

Percentage *not-P & Q* responses:

	Social contract prediction: Social contract > non-SC permission		Permission schema prediction: Social contract = non-SC permission
	high	low	high high
Exp. 8:	65%	> 0%	
Exp. 9:	80%	> 10%	

According to permission schema theory, both rules should have elicited a high percentage of *not-P & Q* responses, because both were switched permission rules in the Rule 3 format. However, the non-SC permission rules, which lacked the cost–benefit structure of a social contract, elicited low percentages of this response (Exp. 8: 0%; Exp. 9: 10%). In contrast, the percentage of *not-P & Q* responses elicited by the switched social contracts was quite high (Exp. 8: 65%; Exp. 9: 80%). The difference in responses between the switched social contracts and the switched non-SC permission rules was large and significant, just as social contract theory predicted it would be (Exp. 8: 65% vs. 0%, $Z = 4.39$, $p < .00001$, $\phi = .69$. Exp. 9: 80% vs. 10%, $Z = 4.45$, $p < .000005$, $\phi = .70$).

But what if subjects were *not* able to read an implicit "may" into "If the precondition has been satisfied, then the action is to be taken"? Since permission schema theory does not distinguish between social contract permission rules and non-social contract permission rules, there is no principled way that a permission schema theorist could take the position that subjects can read a "may" into the switched social contract but not into the switched non-SC permission rule. But let us entertain this hypothesis for a moment anyway.

Subjects should be able to read in the implicit "may" as long as they are able to clearly distinguish a "precondition that is satisfied" from a "condition that has occurred". If they cannot, then they might mistake "If the precondition has been satisfied, then the action is to be taken"—a switched permission rule—for an obligation rule of the form "If condition C occurs, then action A must be taken". Although the response that pragmatic reasoning theory predicts for a switched permission rule is *not-P & Q*, the response it predicts for an obligation rule of this form is *P & not-Q*.

From a social contract point of view, permission and obligation are two sides of the same coin: accepting the benefit obligates you to pay the cost, and paying the cost entitles you to the benefit. These relations hold whether

you accept the benefit before you pay the cost, or after. Not so for pragmatic reasoning theory, however.

In pragmatic reasoning theory, "If you take the benefit, then you pay the cost"—a social contract—can be represented either as a permission ("If you take action A, then you must satisfy precondition P") or as an obligation ("If condition C occurs, then you must take action A"). If the cost must be paid first, then the permission schema is appropriate; if it may be paid after the benefit is accepted, then the obligation schema is appropriate. (Cheng and her colleagues do not say what schema is used if the exchange is simultaneous, or if the timing is left unspecified.) Thus, where social contract theory needs only one algorithm to encompass these 2–4 situations, pragmatic reasoning theory needs two or more. Beyond this consideration of parsimony, however, the over-general nature of the permission and obligation schemas' level of representation presents a theoretical problem. If one cannot tell a "precondition that has been satisfied" from a "condition that has occurred", then one cannot tell a switched permission rule from an obligation rule. "If a student is to be assigned to Grover High School, then that student must live in Grover City" is clearly a permission rule, but what is, "If a student lives in Grover City, then that student is to be assigned to Grover High School"? Is living in Grover City a precondition that is satisfied, or a condition that has occurred? The answer to this question will decide whether assigning the student to Grover High School is an action that *may* be taken, or an action that *must* be taken: whether the rule expresses a switched permission, or an obligation.

This highlights a weakness in pragmatic reasoning theory: its representations are so abstract that it is difficult to distinguish a switched permission rule from an obligation rule, and therefore difficult to tell which of these two interpretations the theory predicts in any particular case. Social contract theory does not have this problem because its representational categories are not so broad, and because it represents the world in terms of *perceived* costs and *perceived* benefits, which are, by definition, distinguishable. The deontic "may" will always be assigned to the benefit clause, and the deontic "must" to the cost clause, no matter what their position in a social contract rule.

So, does pragmatic reasoning theory fare any better if we assume that subjects viewed the non-SC rules as obligation rules rather than as switched permissions? Unfortunately it does not. If subjects had obligation schemas, then an obligation rule would elicit a high percentage of *P & not-Q* responses. However, only 30% of subjects made this response to the non-SC rule in Experiment 8, and only 5% of subjects did in Experiment 9. The difference in percentage of *not-P & Q* answers elicited by the switched social contract and *P & not-Q* answers elicited by the non-SC "obligation rule" is still large

and significant (Exp. 8: 65% vs. 30%, $Z = 2.22$, $p < .015$, $\phi = .35$. Exp. 9: 80% vs. 5%, $Z = 4.80$, $p < .000001$, $\phi = .76$). Thus, even if one wanted to characterize this experiment as one in which a switched social contract was tested against a non-SC obligation rule, the social contract theory prediction would be confirmed, and the pragmatic reasoning theory prediction falsified. Obligation rules that lack the cost–benefit structure of a social contract do not elicit the response predicted by pragmatic reasoning theory.

What if the pragmatic reasoning theorist wanted to say that the non-SC rules activate a permission schema in some subjects and an obligation schema in others (and could find some way of explaining why this does not happen with social contract rules)? Even if one counts both *not-P & Q* and *P & not-Q* as correct for the non-SC rules, only 30% of responses would be consistent with pragmatic reasoning theory in Experiment 8, and 15% in Experiment 9. Moreover, the switched social contracts would still elicit significantly more *not-P & Q* responses than both responses combined for the non-SC rules (Exp. 8: 65% vs. 30%, $Z = 2.22$, $p < .015$, $\phi = .35$. Exp. 9: 80% vs. 15%, $Z = 4.12$, $p < .000025$, $\phi = .65$). So the pragmatic reasoning theorist would still be left with the problem of explaining why cost–benefit representations work, and action–(pre)condition representations do not.

Experiments 5–9 establish that social contract rules elicit high percentages of social contract responses, whereas permission rules that lack the cost–benefit structure of a social contract do not. But do non-SC permission and obligation rules elicit a content effect of any kind? Experiments 5, 6, 8 and 9 tested non-SC permission rules that had social purposes: 30% of subjects (24 out of 80) gave responses consistent with permission schema theory to these problems (just to load the dice in favor of pragmatic reasoning theory, I counted both permission *and* obligation answers as correct for the switched permission rules of Experiments 8 and 9). This figure can be compared to the percentage of falsifying, *P & not-Q* responses to the unfamiliar descriptive problems and the abstract problems of Experiments 1–4 in Part I. (Performance on abstract problems is the normal yardstick for judging content effects in the Wason selection task literature, but the unfamiliar descriptive problems are also an interesting standard as they relate concrete terms rather than letters and numbers.)

In Part I, falsifying responses were elicited from 25% of subjects for the abstract problems (24 out of 96), and from 21% of subjects for the unfamiliar descriptive problems (20 out of 96).[9] By either standard, non-SC permission

[9]Although subjects in Part I were from Harvard, and subjects in Part II were from Stanford, there is no reason to believe that the subject populations from these two highly competitive universities differed in any significant way. Moreover, in an unpublished study testing (among other things) unfamiliar descriptive rules, Stone (1986) found that the falsification rate among Stanford students was also approximately 25%.

rules do not elicit a content effect on the Wason selection task (non-SC permission vs. abstract problem: 30% vs. 25%, $Z = 0.74$, n.s.; non-SC permission vs. unfamiliar descriptive problem: 30% vs. 21%, $Z = 1.39$, n.s.).

Discussion for Part II

By hypothesis, social contract algorithms represent the world at a lower level of abstraction than do permission and obligation schemas. Social contract algorithms represent the world in terms of perceived benefits and perceived costs, whereas permission and obligation schemas represent the world in terms of "actions to be taken" and "(pre)conditions to be satisfied". All social contracts fit the representational structure of a permission rule, but not all permission rules have the cost–benefit structure of a social contract. This means that the permission schema's proposed domain of operation is larger than that of the social contract algorithms.

This hypothesized difference in level of representation has empirical consequences. According to permission schema theory, permission rules that are *not* social contracts should elicit a content effect on the Wason selection task; according to social contract theory, they should not.

This prediction was tested in Experiments 5, 6, 8 and 9, in which permission rules that lacked the cost–benefit structure of a social contract were tested against social contract rules. Because Cheng and Holyoak hypothesized that the provision of a "social purpose" helps activate the permission schema, all non-social contract (non-SC) permission rules were given a social purpose. These experiments constitute critical tests that allow one to decide which kind of representation is psychologically real: the action–precondition representation of permission schema theory, or the cost–benefit representation of social contract theory.

The first critical test compared performance on standard social contracts to performance on non-SC permission rules in the format of Rule 1 of the permission schema (Experiments 5 and 6). According to permission schema theory, both rules should have elicited a high percentage of *P & not-Q* responses; according to social contract theory, only the standard social contract should. In both experiments, the prediction of social contract theory was verified, and the prediction of permission schema theory was falsified. Seventy-five to 80% of subjects chose *P & not-Q* in response to the standard social contract, compared to only 30–45% of subjects for the non-SC permission rule. This result was found whether the rule tested was set in a familiar milieu (Exp. 5), or a culturally alien one (Exp. 6).

The second critical test compared performance on switched social contracts

to performance on switched, non-SC permission rules—rules in the format of Rule 3 of the permission schema (Experiments 8 and 9). According to permission schema theory, both rules should elicit high levels of *not-P & Q* responses; according to social contract theory, only the switched social contracts should. Again, the results verified the prediction of social contract theory and falsified the prediction of permission schema theory. Sixty-five to 80% of subjects chose *not-P & Q* for the switched social contracts, compared to 0–10% of subjects for the switched, non-SC permission rules. Even if one assumes that some subjects interpreted the non-SC problem as an obligation rule, rather than as a switched permission rule, the percentage of responses consistent with pragmatic reasoning theory (switched permission: *not-P & Q*; obligation: *P & not-Q*) rises only to 15–30%. Again, this result held whether the rule tested was set in a familiar milieu (Exp. 8), or a culturally alien one (Exp. 9).

Permission schema theory maintains that the provision of a social purpose enhances performance on permission rules; social contract theory maintains that social purposes do not enhance performance unless they happen to confer the cost–benefit structure of a social contract on the rule. Performance on a non-SC permission rule that had a social purpose (Exp. 5) was compared to performance on a non-SC permission rule that lacked a social purpose, but was otherwise identical (Exp. 7). There was no significant difference in performance between the rule that had a social purpose and the rule that lacked one. This supports the argument made earlier, that the only reason Cheng and Holyoak (1985) found facilitation for rules with a social purpose was because the social purposes they happened to choose conferred the cost–benefit structure of a social contract on the rules they tested. It also suggests that the poor performance found in the previous literature for non-SC permission rules was due to their lack of a proper cost–benefit structure, and not due to their lack of a social purpose.

Finally, the results of these experiments indicate that non-SC permission rules do not elicit a content effect of any kind. The percentage of responses consistent with pragmatic reasoning theory that were elicited by non-SC permission rules was not significantly different than the percentage of falsifying responses elicited by either the abstract problems or the unfamiliar descriptive problems tested in Part I.

The results of each of the experiments in Part II verify the predictions of social contract theory, and falsify the predictions of permission schema theory. They indicate that the cost–benefit representations of social contract theory have psychological reality, whereas the action–precondition representations of permission schema theory do not. In so doing, they establish that social contract theory posited the correct domain of operation: the proposed

"permission schema effect" is not found for all permission rules, but only for those that have the cost–benefit structure of a social contract. This indicates that the action–precondition representations of permission schema theory are, indeed, over-general.

To be viable, what standards must be met by a domain-general theory of the content effect on the Wason selection task?

Applying the Wason selection task literature to the issue of modularity in human reasoning is sure to remain controversial, and to provoke questions concerning the soundness of the interpretation advanced here. To be responsible, however, such alternative interpretations as might be advanced need to meet the empirical and logical standards already met by social contract theory.

Obviously, in addition to retrodicting the content effects—and non-effects—already established in the literature, any viable theory must make the same predictions as are supported in this study, that is, they must: (1) predict successful operation on unfamiliar materials; (2) predict a procedure that is isomorphic with the "look for cheaters" procedure; and (3) operate on the "social contract" problem domain, *as well as on whatever other content domains its theory specifies.* At present, no published theory meets these requirements.

Although it is difficult to second-guess human ingenuity, it seems likely that the only kind of theory many cognitive scientists see as potentially capable of meeting these criteria is some variant of pragmatic reasoning theory, which accepts domain-specific procedures, but explains them as the product of an overarching domain-general process. Given the data presented in this article, both the availability theorist and the pragmatic reasoning theorist might agree with social contract theory's claims about the structure of the algorithms that guide reasoning about social exchange, but they might disagree with its claims about the origin of these algorithms. They might argue that the social contract algorithms were produced by, for example, a general-purpose inductive process.

In psychology, the term "induction" does not refer to any particular set of known algorithms. Instead, it refers to the cognitive processes—as yet unspecified—that allow us to abstract general principles from finite sets of data. Thus, before the pragmatic reasoning approach can be treated empirically as a theory about induction, it must be turned into a *specified* theory: it must lay out mechanistically specified domain-general procedures that take modern human experience as it is statistically encountered as input, and produce social contract algorithms as output. At present, no such theory exists (see,

e.g., Holland, Holyoak, Nisbett & Thagard, 1986, p. 2). The connection made by Cheng and her colleagues between their structurally well-specified permission, obligation, causal and covariation schemas, and the hypothetical domain-general inductive process that is supposed to have produced them, is, at this time, simply an assertion without content. It is possible to test the structure of their schemas, but they have no published theory or data demonstrating that these schemas could have arisen through a general-purpose inductive process. At present, therefore, their assertion of such a relationship cannot be falsified or experimentally evaluated.

Once such a theory is advanced, it must pass the following test as well: from the statistical distribution of *modern* human experience, it must predict which domains of human activity will have associated pragmatic reasoning schemas. A Darwinian approach does predict the distribution of domain-specific procedures, and does so through assessing the importance and recurrence of specific problem domains in our species' evolutionary history, and not on the basis of modern experience. While social exchange was a crucial adaptation for hunter–gatherers, permission from "institutional authorities" was not, so an evolutionary approach predicts the existence of social contract algorithms but not the existence of permission algorithms. Furthermore, which production rules a schema has should be deducible from the theory: it is not good enough to infer a modus tollens production rule (such as Rule 4 of the permission schema) by observing, post hoc, whether or not subjects falsify in a given domain.

Although no specified theory yet exists, by making some minimal assumptions about what explanatory variables would be most relevant to an inductive process that is domain-general,[10] we can ask whether it is *plausible* to believe that such a process produced the social contract algorithms.

Ever since the British empiricists argued that the experience of spatially and temporally contiguous events is what allows us to jump from the particular to the general, from sensations to objects, from objects to concepts, *recurrent experience* has been seen as the engine that drives both associationism and induction. The sight of one white swan may mean nothing, but if one sees a hundred white swans and no black ones, one might begin to associate whiteness with swans, and one might induce the rule "all swans are white".

[10]Note: no one doubts that induction occurs, or that it can produce useful generalizations, including schematic rules, if it operates over a suitably constrained set of inputs. In fact, Darwinian algorithms can be thought of as bundles of constraints that organize experience into adaptively meaningful categories, schemas and frames, thereby making induction possible. The question addressed here is whether the pattern of results found could have been produced by an inductive process that does *not* operate on experience that is pre-structured by Darwinian algorithms.

The more experiences one has, the stronger the association becomes, and the more likely one is to induce a rule.

Let us say, then, that individuals can be shown to have abstract schemas for detecting violations of rules from category A, but not for detecting violations of rules from category B. To account for such a result, one must invoke differential experience: one must show that individuals have encountered category A rules more frequently than category B rules. If differential performance does not correlate with differential experience, then it is unlikely that the schema was produced by a domain-general inductive process.

We can therefore ask, can the presence of social contract algorithms be explained by a general-purpose theory that invokes differential experience as its major explanatory variable? There are four reasons why this is implausible:

(1) The experiments of Part II show that social contract algorithms represent the world in terms of costs and benefits, rather than in terms of actions and preconditions. But why would a domain-general inductive process stop at the cost–benefit level of abstraction, rather than continue "up" to create the more abstract action–precondition representations originally postulated by pragmatic reasoning theory? Differential experience cannot explain this. Social contracts are a *subset* of all permission rules, so, by definition, people have had *more* experience relating actions to preconditions than they have had relating costs to benefits. And it is not the case that people rarely encounter non-SC permission rules in real life. Daily we are beset by a plethora of bureaucratic rules, institutional rules, traffic rules, etiquette rules, rules set by employers, parents, teachers, and so on, that either have no cost–benefit structure, or at the very least, for which the cost–benefit structure is not readily apparent to us. Nor can "pragmatics" explain why the inductive process is arrested at the cost–benefit level, because surely people have the goal of either obeying, or disobeying, these rules and regulations. In fact, permission rules are Cheng and Holyoak's paradigmatic case of a pragmatically useful rule.

Differential experience would therefore lead one to expect the action–precondition schemas that Cheng and Holyoak originally predicted. Yet the results of these experiments run counter to this prediction. We have schemas corresponding to the less common, rather than to the more common, category of rule.

(2) If induction had produced abstract schemas for reliably detecting violations of social contracts, then it should also have produced abstract schemas for reliably detecting violations of rules from other commonly encountered categories. Yet people do not appear to have such schemas

for a number of categories that are, arguably, far more common than social contracts. Subjects are not good at detecting violations of descriptive rules (called "covariation" rules by Cheng & Holyoak) or of non-SC permission rules, and, although the final word is not in yet, preliminary evidence suggests that subjects are also not good at detecting violations of causal rules (e.g., the electron repulsion problem tested by Cheng et al., 1986; the chili pepper problem tested by Cosmides, 1985, p. 253; Stone, 1986; Stone & Cosmides, in prep.).

Although no one knows the exact statistical frequency with which most people encounter rules from these various categories, common sense tells us that they are all very common, and exactly the sort of relations that a general-purpose inductive process should build schemas for reasoning about. Descriptive relations, for example, are ubiquitous—"the sky is blue", "the restaurant is around the corner", "this water is polluted", "the meat is spoiled"—they are the relations people use to describe and act on the world. Every declarative sentence contains one or more descriptive relation. They are probably far more common than social contracts; at the very least, there is no reason to believe that descriptive rules are *less* common than social contract rules. Moreover, many of the descriptive relations we hear of or think of are false ("the check is in the mail", "blondes have more fun"), and even those that are more or less true are frequently violated: "the sky is blue" is violated every gray and stormy day. The same is true of causal rules. Many of the causal rules we hear are false, from old wives' tales like "celery makes your hair curly" to political analyses of what is causing the most recent economic recession. We test hypothesized causal relations in making repairs, in trying variations on recipes, in trying to understand and predict the vicissitudes of our world. We have animated discussions—about society, politics, religion—when our friends' explanations of "how things work" do not jibe with our own personal experiences.

Furthermore, if we do learn about the world via a general-purpose inductive process, then we must experience *many* violations of descriptive and causal rules: at least one for each incorrect hypothesis our induction mill generates. After all, induction does not generate all and only correct hypotheses. And injecting "pragmatics" into the discussion does not solve anything. The search for knowledge about how the world is, and how it works, *is* a goal-driven activity, and it is hard to imagine solving any problem without the ability to detect whether the descriptive or causal knowledge relevant to solving it is true or false. Thus it would be very useful to be good to detecting violations of descriptive and causal rules, and "pragmatic contexts" would have provided ample opportunity to test such rules.

Similarly, from the youngest age we are inundated with non-SC permission rules in school, at home, at work, and in religious institutions, as discussed above. Assuming there were some reason why the inductive machine would not categorize SC permission rules *with* non-SC permission rules, why would it build a schema that allows us to detect violations of SC permission rules, but not build a separate one that allows us to detect when non-SC permission rules have been disobeyed?

In short, the results indicate that subjects have abstract schemas for reliably detecting violations of social contract rules, but they do not have abstract schemas for reliably detecting violations of other commonly encountered rules: descriptive rules, non-SC permission rules and, possibly, causal rules. A general-purpose cognitive process would have produced schemas for detecting violations of all of them, or none of them.

(3) The fact that subjects choose the *P* card, but fail to choose the *not-Q* card, on the Wason selection task led Wason and Johnson-Laird (1972) to argue that subjects try to verify, rather than falsify, conditional rules; indeed, in permission schema theory, these two card choices are made by two entirely different production rules (Rules 1 and 4). Analogously, it is not clear why induction would construct a production rule that makes individuals so good at detecting *violations* of a social contract, rather than simply creating rules that look for compliance with one. In the real world, compliance is the rule, violation the exception: every time a store lets you walk out with the goods you have paid for, you have experienced compliance with a social contract. Differential experience favors compliance, therefore one would expect a general-purpose inductive machine to create schemas that look for compliance, not cheating. The permission schema, for example, would have Rule 1, but lack Rule 4.

At best, a subject's ratio of compliance-to-cheating episodes should be the idiosyncratic product of different life experiences, and, therefore, the number of individuals who possess a production rule that makes them good at looking for cheaters—as evidenced by choosing the "cost not paid" card—should vary substantially. Yet the vast majority of subjects chose this card in all experiments testing social contract rules. With induction, invariance in experience is required to produce invariance in production rules. There is no compelling reason to believe that episodes of cheating during an individual's early life are sufficiently frequent, or that the frequent experience of such episodes is sufficiently universal, for induction to invariantly insert a "look for cheaters" production rule into our social contract algorithms.

(4) The notion that one can learn what constitutes cheating by experiencing violations of a social contract is highly problematic. You cannot know that a violation of a social contract has occurred unless you already know what counts as a violation. But this is the very thing that induction is supposed to teach you!

Similarly, trial and error learning requires some, non-learned, definition of error—you cannot learn a definition that must already exist for that learning to occur in the first place. A general-purpose, content-independent learning mechanism needs a general-purpose, content-independent definition of error. Logical falsification, for example, is a content-independent definition of error or violation. But the definition of violation for social contracts is quite specific: cheating is defined as absconding with a benefit when you have not paid the required cost. It conforms to no known content-independent definition of error; it certainly does not map onto logical falsification, as a consideration of the *not-P & Q* response to switched social contracts demonstrates. Without built-in, domain-specific knowledge defining what counts as cheating, how could one develop a "look for cheaters" procedure?

An evolutionarily based social contract theory handles the above four issues with ease:

(1) Our social contract algorithms operate on cost–benefit representations (rather than action–precondition ones) because this is the highest level of abstraction at which social exchange can be understood.
(2) We have social contract algorithms because social exchange is an evolutionarily crucial domain. Other problems tested so far have not tapped into evolutionarily important domains of human activity; subjects have no specialized procedures causing them to choose the *not-Q* card (or any other card) for such problems, so they usually don't. In solving these problems, subjects use general knowledge of the world, of contingency, or of syntax.
(3) Procedures that make us very good at detecting cheating were directly programmed into our Darwinian algorithms for reasoning about social exchange because the failure to detect cheating results in large fitness costs. The capacity to engage in social exchange could not have evolved in the first place unless we had such procedures.
(4) The problem of learning the definition of cheating by trial and error does not arise because the correct definition is directly programmed into our social contract algorithms.

Social contract theory not only provides the most parsimonious explanation of the data, but the assumption that some innate algorithms are special-pur-

pose and content-dependent is also more parsimonious from the standpoint of evolutionary theory. Social exchange is a domain for which the evolutionarily predicted computational theory is complex, and the fitness costs associated with "errors" are large. Even if it were possible for a domain-general information-processing strategy to construct social contract algorithms—and it is by no means clear that it *is* possible—it is not reasonable to expect that natural selection would leave learning in such a domain to the vagaries of a general-purpose mechanism that builds schemas or not, depending on the vicissitudes of idiosyncratic personal experience. Successfully conducted social exchange was such an important and recurrent feature of hominid evolution that a reliable, efficient cognitive capacity specialized for reasoning about social exchange would quickly be selected for. A general-purpose learning mechanism would either be supplanted or be used only for learning in other domains.

Conclusions

Whether the human cognitive architecture contains an array of special-purpose, domain-specific, procedure-rich modules, or consists entirely of a few, major, domain-general information-processing mechanisms, is very much at issue in modern cognitive science. To date, this debate (involving such issues as learnability, innateness, and so on) has been conducted primarily within the field of psycholinguistics, and has left most other subfields of cognition largely untouched. Human reasoning, especially, has been considered quintessentially domain-general: the innate processes hypothesized—whether "logical", "inductive" or associationistic—have been thought of as operating consistently, regardless of content, with content-dependent performance attributed to differential experience. If this and other empirical studies establish that even human reasoning is not unitary and domain-general, but instead governed by an array of special-purpose mechanisms, this will provide substantial support for a modular approach to cognitive psychology. These studies have been designed to contribute to the resolution of this issue, by widening the debate from psycholinguistics into the field of human reasoning, and they provide empirical support for an evolutionary and modular approach outside of psycholinguistics.

Social contract theory started with the hypothesis that: (1) humans have algorithms specialized for reasoning about social exchange; (2) these algorithms will have certain structural properties, predicted by natural selection theory; and (3) these algorithms are innate, or else the product of experience structured by innate algorithms that are specialized for reasoning about social

exchange. The evidence presented in this article supports the first two propositions, and lends substantial plausibility to the third (see Discussion for Part II). The subjects tested were not adept at looking for violations of descriptive rules, or of permission rules that lacked the cost–benefit structure of a social contract, but when they were reasoning about conditional rules that did have the cost–benefit structure of a social contract, they consistently "looked for cheaters". The cards that represented potential cheaters (the "benefit accepted" card and the "cost not paid" card) were chosen no matter what logical category they corresponded to, and no matter how unfamiliar the social contract rule. Specifically, three dimensions were manipulated experimentally to allow critical tests to be made between social contract theory and other, competing theories:

(1) *Unfamiliarity:* Because innate mechanisms hypothesized to function as frame-builders must be able to act on unfamiliar experience, it is a prediction of social contract theory that the social contract algorithms will function on unfamiliar materials. For this reason, extreme unfamiliarity was used, including not just unfamiliarity of rule, but unfamiliarity of element, context and relationship. The effort was to provide materials as unfamiliar as possible while still allowing the criterially necessary cost–benefit structure to be recoverable to the reader. The manipulation of unfamiliar elements was especially useful in falsifying the availability theories of reasoning. No matter how unfamiliar the social contract rule tested, subjects chose the "cost not paid" card and the "benefit accepted card"—a result that the availability theories can neither predict nor explain.

(2) *Structure of the invoked procedure:* Social contract theory made predictions about the structure of the algorithm, indicating that it should include a "look for cheaters" procedure that focuses attention specifically on individuals who have accepted a benefit, and individuals who have not paid the cost. Subjects were just as likely to choose the "benefit accepted" card and the "cost not paid" card when these corresponded to the illogical *not-P & Q* response, as when they corresponded to the logically falsifying *P & not-Q* response. Finding performance consistent with this prediction concerning the structure of the procedure, regardless of which logical category the "cost not paid" and "benefit accepted" cards corresponded to, was especially useful in falsifying the hypothesis that social contract problem content facilitates use of the propositional calculus, or promotes "logical" reasoning per se.

(3) *Level of representation/domain of operation:* Social contract theory made the prediction that the "look for cheaters" procedure would be invoked

with increasing likelihood the more clearly the problem content was identifiable as a "social contract" problem, that is, the more clearly the subject could represent the situation as one in which desired benefits were rationed by costs to be paid. In contrast, permission schemas were hypothesized to represent the world in terms of "actions to be taken" and "preconditions to be satisfied". Manipulating the level of representation to identify the hypothesized domain of operation was particularly useful in falsifying pragmatic reasoning theory's hypothesis that "permission content" with a "social purpose" invokes a permission schema. This hypothesis was falsified because (a) the "look for cheaters" procedure was invoked only by "social contract" content, and (b) the permission schema hypothesized to exist was not invoked by permission content when that content was not also social contract content.

On an empirical level, then, the only candidate theory to meet these experimental tests is social contract theory. The hypothesis that humans have social contract algorithms predicts and explains the results of all nine experiments. Moreover, social contract theory explains the apparently contradictory literature attempting to stalk the "elusive" content effect on the Wason selection task: robust and replicable content effects are found only for rules that are standard social contracts—the only rules for which the predicted social contract response is also the logically falsifying response.

Moreover, because the computational theory of social exchange was developed independent of the Wason selection task, it entails many other testable predictions (Cosmides, 1985; Cosmides & Tooby, 1989). The contractual conditions implicit in social contract interactions are analogous to, but more detailed than, the logical and modal relations expressed by the permission schema's four production rules. These include concepts of obligation, entitlement and intentional causality, in addition to conditions of restitution or punishment for cheating. These contractual conditions predict richly structured trains of inference that should guide attention, memory and decision-making in this domain. Engaging in social exchange is an extremely complex computational feat; it seems easy for the same reason that seeing does: we have Darwinian algorithms that are up to the task.

The finding that adult subjects are very adept at detecting potential "cheaters" on a social contract, even when it is unfamiliar and culturally alien, stands in marked contrast to the repeated finding that they are not skilled at detecting the potential invalidity of descriptive or permission rules, familiar or unfamiliar. The ontogeny of the algorithms that produce these results remains an open question. It is possible that they are, in some carefully delimited sense, learned. However, the mental processes involved appear to

be powerfully structured for social contracts, yet weakly structured for other elements and relations drawn from common experience. This implies that the learning process involved is guided and structured by special-purpose innate algorithms, just as learning a natural language is guided and structured by the innate algorithms of a language acquisition device.

Thus, although two series of experiments cannot decide the nature of the human cognitive architecture, they do lend empirical credence to the modular view. More importantly, however, they can function as a case study of how evolutionary biology can contribute to the study of human information-processing mechanisms. If cognitive psychologists use evolutionary biology to develop computational theories of adaptive information-processing problems, they will have at their disposal a powerful new set of tools for investigating the design features of human information-processing mechanisms, tools which complement existing methodology in cognition.

Appendix: Texts of problems

(1) **Unfamiliar standard social contract—Social law:** Used in Experiments 1, 3, 6 and 9. For Experiments 3 and 9 the same text was used, but the rule was "switched".

You are a Kaluame, a member of a Polynesian culture found only on Maku Island in the Pacific. The Kaluame have many strict laws which must be enforced, and the elders have entrusted you with enforcing them. To fail would disgrace you and your family.

Among the Kaluame, when a man marries, he gets a tattoo on his face; only married men have tattoos on their faces. A facial tattoo means that a man is married, an unmarked face means that a man is a bachelor.

Cassava root is a powerful aphrodisiac—it makes the man who eats it irresistible to women. Moreover, it is delicious and nutritious—and very scarce.

Unlike cassava root, molo nuts are very common, but they are poor eating—molo nuts taste bad, they are not very nutritious, and they have no other interesting "medicinal" properties.

Although everyone craves cassava root, eating it is a privilege that your people closely ration. You are a very sensual people, even without the aphrodisiacal properties of cassava root, but you have very strict sexual mores. The elders strongly disapprove of sexual relations between unmarried people, and particularly distrust the motives and intentions of bachelors.

Therefore, the elders have made laws governing rationing privileges. The one you have been entrusted to enforce is as follows:

"If a man eats cassava root, then he must have a tattoo on his face."

Cassava root is so powerful an aphrodisiac, that many men are tempted to cheat on this law whenever the elders are not looking. The cards below have information about four young Kaluame men sitting in a temporary camp; there are no elders around. A tray filled with cassava root and molo nuts has just been left for them. Each card represents one man. One side of a card tells which food a man is eating, and the other side of the card tells whether or not the man has a tattoo on his face.

Your job is to catch men whose sexual desires might tempt them to break the law—if any get past you, you and your family will be disgraced. Indicate only those card(s) you definitely need to turn over **to see if any of these Kaluame men are breaking the law.**

The four cards read: "eats cassava root", "no tattoo", "eats molo nuts", "tattoo".

(2) **Unfamiliar standard social contract—Private exchange:** Used in Experiments 2 and 4. For Experiment 4 the same text was used, but the rule was "switched".

You are an anthropologist studying the Kaluame, a Polynesian people who live in small, warring bands on Maku Island in the Pacific. You are interested in how Kaluame "big men"—chieftans—wield power.

"Big Kiku" is a Kaluame big man who is known for his ruthlessness. As a sign of loyalty, he makes his own "subjects" put a tattoo on their face. Members of other Kaluame bands never have facial tattoos. Big Kiku has made so many enemies in other Kaluame bands, that being caught in another village with a facial tattoo is, quite literally, the kiss of death.

Four men from different bands stumble into Big Kiku's village, starving and desperate. They have been kicked out of their respective villages for various misdeeds, and have come to Big Kiku because they need food badly. Big Kiku offers each of them the following deal:

"If you get a tattoo on your face, then I'll give you cassava root."

Cassava root is a very sustaining food which Big Kiku's people cultivate. The four men are very hungry, so they agree to Big Kiku's deal. Big Kiku says that the tattoos must be in place tonight, but that the cassava root will not be available until the following morning.

You learn that Big Kiku hates some of these men for betraying him to his enemies. You suspect he will cheat and betray some of them. Thus, this is a perfect opportunity for you to see first hand how Big Kiku wields his power. The cards below have information about the fates of the four men. Each card

represents one man. One side of a card tells whether or not the man went through with the facial tattoo that evening and the other side of the card tells whether or not Big Kiku gave that man cassava root the next day.

Did Big Kiku get away with cheating any of these four men? Indicate only those card(s) you definitely need to turn over **to see if Big Kiku has broken his word to any of these four men.**

The cards read: "got the tattoo", "Big Kiku gave him nothing", "no tattoo", "Big Kiku gave him cassava root".

(3) **Unfamiliar descriptive problem:** Used in Experiments 1–4. The rule used matched that used for the corresponding social contract problem (standard or switched).

You are an anthropologist studying the Kaluame people, a Polynesian culture found only on Maku Island in the Pacific. Before leaving for Maku Island you read a report that says some Kaluame men have tattoos on their faces, and that they eat either cassava root or molo nuts, but not both. The author of the report, who did not speak the language, said the following relation seemed to hold:

"If a man eats cassava root, then he must have a tattoo on his face."

You decide to investigate your colleague's peculiar claim. When you arrive on Maku Island, you learn that cassava root is a starchy staple food found on the south end of the island. Molo nuts are very high in protein, and grow on molo trees, which are primarily found on the island's north shore.

You also learn that bachelors live primarily on the north shore, but when men marry, they usually move to the south end of the island. When a Kaluame man marries, he gets a tattoo on his face; only married men have tattoos on their faces. A facial tattoo means that a man is married, an unmarked face means that a man is a bachelor. Perhaps men are simply eating foods which are most available to them.

The cards below have information about four Kaluame men sitting in a temporary camp at the center of the island. Each man is eating either cassava root or molo nuts which he has brought with him from home. Each card represents one man. One side of a card tells which food a man is eating and the other side of the card tells whether or not the man has a tattoo on his face.

The rule laid out by your colleague may not be true; you want to see for yourself. Indicate only those card(s) you definitely need to turn over **to see if any of these Kaluame men are breaking the rule.**

The cards read: "no tattoo", "tattoo", "eats cassava root", "eats molo nuts".

(4) **Unfamiliar standard social contract—Social law:** Used in Experiments 1, 3, 6 and 9. For Experiments 3 and 9 the same text was used, but the rule was "switched".

You are an anthropologist studying the Namka, a hunter–gatherer culture living in the deserts of southwest Africa. You are particularly interested in whether Namka boys obey the laws of their people.

Every full moon there is a special feast in which a duiker—a small antelope—is slaughtered and eaten. Duiker meat is quite scarce and delicious—a real treat. Eating duiker meat is a privilege that must be earned.

For boys, this privilege is governed by the following law:

"If you eat duiker meat, then you have found an ostrich eggshell."

Finding ostrich eggshells is a sophisticated and difficult task which takes a boy years to learn. Having found an ostrich eggshell on your own is therefore a sign that you have mastered the most difficult skills of hunting. For the Namka, it represents a boy's transition into manhood.

You wonder if Namka boys cheat on this law when nobody is looking. You decide to hide behind some bushes and watch. During the course of the feast of the full moon, you see four different boys approach the roasted duiker while no one else is looking.

The cards below have information about these four boys. Each card represents one boy. One side of a card tells whether a boy has ever found an ostrich eggshell, and the other side of the card tells whether that boy took any of the roasted duiker meat.

The smell of the roasting duiker is truly tempting to the boys. You want to know if any of them cheated on the law. Indicate only those card(s) you definitely need to turn over **to see if any of these boys have broken the law.**

The four cards read: "eats some duiker meat", "has never found an ostrich eggshell", "does not eat any duiker meat", "has found an ostrich eggshell".

(5) **Unfamiliar standard social contract—Private exchange:** Used in Experiments 2 and 4. For Experiment 4 the same text was used, but the rule was "switched".

The Namka are a hunter–gatherer people who live in small bands in the deserts of southwest Africa. You are an anthropologist interested in whether members of different Namka bands can trust each other.

Bo is a crafty old Namka man in the band you are studying. He is always accidentally breaking his ostrich eggshell and would like to "stockpile" some—the Namka use ostrich eggshells as canteens because they are light

and hold lots of water. He sees his opportunity when four men from a neighboring band stumble into camp one morning.

The four men have been on a long and unsuccessful hunting expedition. They are hungry, and they want to be able to bring meat back to their families. Bo approaches each man privately and offers him the following deal:

"If you give me your ostrich eggshell, then I'll give you duiker meat."

Bo explains that his wife is skinning the duikers today, and they won't be ready until tomorrow. However, he will need the eggshell by this evening for his son, who is leaving tonight on a week long hunting expedition. Each man accepts Bo's offer, and agrees to meet him alone in a secluded spot tomorrow to consummate the deal.

You find this deal interesting, because you happen to know that Bo, who is a rather unscrupulous character to begin with, has very little duiker meat and a large family to feed. It is perfectly possible that he will cheat some of these men. You decide to "spy" on Bo and see.

The cards below have information about the four deals Bo made with these four men. What happened in one deal had no effect on the outcome of any other deal. Each card represents one man. One side of a card tells whether or not the man gave his ostrich eggshell to Bo that evening, and the other side of the card tells whether or not Bo gave that man duiker meat the next day.

Did Bo get away with cheating any of these four men? Indicate only those card(s) you definitely need to turn over **to see if Bo has broken his word to any of these four men.**

The cards read: "He gave his ostrich eggshell to Bo", "Bo gave him nothing", "He gave Bo nothing", "Bo gave him duiker meat".

(6) **Unfamiliar descriptive problem:** Used in Experiments 1–4. The rules used matched those used for the corresponding social contract problem (standard or switched).

You are an anthropologist studying the Namka, a hunter–gatherer culture in the deserts of southwest Africa. Over and over again, you hear various Namka repeat the following saying:

"If you eat duiker meat, then you have found an ostrich eggshell."

Duikers are small antelopes found in the eastern part of the Namka's home range. Both duiker meat and ostrich eggshells are sought by the Namka: they eat the meat and they use the eggshells as canteens because they are light and hold lots of water. Furthermore, duikers frequently feed on ostrich eggs.

As an anthropologist, you don't know if this saying is metaphorical, referring, for example, to clan territories or ritual practices, or if the saying reflects a real relationship the Namka use to guide their foraging behavior. Does it mean that if you find the first you find the second? This is what you are trying to find out.

Is it fact or folklore? Do the Namka *mean* eggshells and duiker meat, or are these things merely symbols for something else entirely? Unfortunately, you don't know their language well enough to ask them. So you decide to investigate whether the rule stated in this saying has any *factual* basis.

Many species of birds populate the area, and in your wanderings you have come across several caches of eggs of various sorts. The cards below have information about four different locations with egg caches. Each card represents one location, and each location has the tracks of one mammal associated with it. One side of a card tells what kind of eggshell you found at a location, and the other side of the card tells which mammal's tracks you found there.

Perhaps the Namka's saying has no factual basis. Indicate only those card(s) you definitely need to turn over **to see if your finds at any of these locations violates the rule expressed in the Namka's saying.**

The cards read: "quail eggshell", "ostrich eggshell", "duiker", "weasel".

(7) **Familiar descriptive problem:** Used in Experiments 1–4. Each experiment used four different versions of the rule: Boston–subway, Boston–cab, Arlington–subway, Arlington–cab.

Part of your new job for the City of Cambridge is to study the demographics of transportation. You read a previously done report on the habits of Cambridge residents which says:

"If a person goes into Boston, then he takes the subway."

The cards below have information about four Cambridge residents. Each card represents one person. One side of a card tells where a person went and the other side of the card tells how that person got there.

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

The cards read: "subway", "Arlington", "cab", "Boston".

(8) **Abstract problem:** Used in Experiments 1–4. The text of this problem is shown in Figure 1. Each experiment used two different versions of the rule: D-3 and C-2.

(9) **Standard social contract—School problem:** Used in Experiments 5 and 8. For Experiment 8 the same text was used, but the rule was

"switched". Another version of the problem used the terms: "Milton High School", "town of Milton", "Crandon High School" and "Crandon City", and, instead of "Belmont", the third town mentioned in the text was "Appleton".

You supervise four women who volunteered to help out at the local Board of Education. When you came into work today, you found the place a-buzz with rumor and innuendo. Your volunteers were supposed to follow certain rules for assigning students from various towns to the appropriate school district. Each volunteer is the mother of a teenager who is about to enter high school, and each processed her own child's documents. So now rumors are flying that your volunteers cheated on the rules when it came to assigning their own children to a school. Here is the situation:

Students are to be assigned either to Grover High School, which is located in Grover City, or to Hanover High School, which is located in the town of Hanover. Grover High is a great school with an excellent record for getting students placed in good colleges. In contrast, Hanover High is a mediocre school with poor teachers and decrepit facilities.

The reason the schools are so different is how willing the parents of each community are to financially support their schools through taxes. Although both communities are equally prosperous, the parents in Grover City have always cared about the quality of their schools, including Grover High, and have been willing to pay for it. In contrast, the parents in the neighboring towns of Hanover and Belmont have never wanted to spend the money, and have opposed any taxes to improve Hanover High.

The Board of Education took these factors into account when it created rules to determine which school a student is to be assigned to; the most important of these rules is:

"If a student is to be assigned to Grover High School, then that student must live in Grover City."

Your volunteers were supposed to follow this rule when processing *all* student documents—including the documents of their own children! You must find out if the rumors are true: did any of your volunteers cheat on this rule when it came to processing their own children's documents?

The cards below have information about the documents of the four volunteers' children. Each card represents the child of one volunteer. One side of a card tells what school the volunteer assigned her son or daughter to, and the other side of the card tells what town that student lives in.

Most parents want their children to get the best education possible, however, some are not willing to pay for it. It is easy to imagine that your

volunteers, being ambitious mothers, might have been tempted to cheat on the rule. Indicate only those card(s) you definitely need to turn over **to see if the documents of any of these students violate the rule.**

The cards read: "Grover High School", "town of Hanover", "Hanover High School", "Grover City".

(10) **Non-social contract permission rule—School problem:** Used in Experiments 5 and 8. For Experiment 8 the same text was used, but the rule was "switched". Another version of the problem used the terms: "Milton High School", "town of Milton", "Crandon High School" and "Crandon City", and, instead of "Belmont", the third town mentioned in the text was "Appleton".

The secretary you replaced at the local Board of Education may have made some mistakes when she processed student documents. It is important that certain rules for assigning students from various towns to the appropriate school district be followed, because the population statistics they provide allow the Board of Education to decide how many teachers need to be assigned to each school. If these rules are not followed, some schools could end up with too many teachers, and other schools with too few.

Students are to be assigned either to Grover High School or to Hanover High School.

Some students live in the town of Grover City, some live in Hanover, and some live in Belmont. There are rules that determine which school a student is to be assigned to; the most important of these rules is:

"If a student is to be assigned to Grover High School, then that student must live in Grover City."

Shortly before she retired, the secretary you replaced was supposed to sort through the documents that specify what town the students' live in, and make school assignments according to this rule. She was a sweet little old lady who had become rather absent-minded, and who often made mistakes when categorizing student documents.

The cards below have information about the documents of four students. Each card represents one student. One side of a card tells what school the retired secretary assigned the student to, and the other side of the card tells what town that student lives in.

You suspect the retired secretary may have inadvertently categorized some of the students' documents incorrectly, so you decide to see for yourself whether she ever violated the rule. Indicate only those card(s) you definitely need to turn over **to see if the documents of any of these students violate the rule.**

The cards read: "Grover High School", "town of Hanover", "Hanover High School", "Grover City".

(11) **No-purpose non-social contract permission rule—School problem:** Used in Experiment 7. The text was identical to that of the non-social contract permission rule listed in (10), except the social purpose was omitted from the first paragraph. The first paragraph therefore read:

The secretary you replaced at the local Board of Education may have made some mistakes when she processed student documents. She had been asked to follow certain rules for assigning students from various towns to the appropriate school district.

(12) **Non-social contract permission rule—Fictitious culture:** Used in Experiments 6 and 9. For Experiment 9 the same text was used, but the rule was "switched".

You are an anthropologist studying the Kaluame people, a Polynesian culture found only on Maku Island in the Pacific. The Kaluame people are divided into two great clans: the Napali, who distinguish themselves by getting tattoos on their faces when they are children, and the Kaloi, who have no facial tattoos. Members of the Napali and Kaloi clans live together in peace and friendship. Important matters are decided by a group of Kaluame elders, half of whom are Napali, and the other half, Kaloi.

Twenty years ago, when you first started studying the Kaluame, the elders became concerned about Maku Island's dwindling resources. The Kaluame do not cultivate food. Instead, they gather food that grows wild. They have two staple foods, which are equally tasty and nutritious: cassava root, a tubor which can grow only on the south end of Maku Island, and molo nuts, from molo trees which can grow only on the island's north shore. The problem is that both cassava root and molo nuts are in short supply, because the Kaluame population has been growing. If all the Kaluame lived on the south end of Maku Island, they would surely exhaust the supply of cassava root; if they all lived on Maku Island's north shore, they would surely exhaust the supply of molo nuts. The elders want their people to live in a balance with nature; they do not want to cause the extinction of either source of food, as this could eventually lead to the extinction of the Kaluame themselves.

Therefore, the elders decided to divide the Kaluame people in half, so that, roughly, one clan would live where the cassava root grows, and eat only cassava root, and one clan would live where the molo nuts grow, and eat only molo nuts. That way, neither food source would be overwhelmed by too many people, and both clans would be well nourished. Everyone agreed that this was a good plan. The only problem was that one clan had more people

than the other, so 10% of the larger clan were asked to live and eat with the smaller clan. They gladly agreed.

The elders expressed the law governing eating arrangements thus:

"If a man eats cassava root, then he must have a tattoo on his face."

When you had left Maku Island, everyone was happily observing this law, hoping that the cassava plants and molo nut trees would flourish as a result, so the next generation would not have to worry about such things.

That was 20 years ago. Now you are returning to Maku Island to continue your study of the Kaluame, and you wonder whether the plan worked. Because the Kaluame are such law-abiding people, the best way to see if the law is still in effect is to watch what the Kaluame are eating; if any of them are breaking it, then it must be because the plants flourished and the elders repealed the law.

The cards below have information about four Kaluame men sitting in a temporary camp at the center of the island. Each man is eating either cassava root or molo nuts, which he brought with him from home. Each card represents one man. One side of a card tells which food a man is eating, and the other side of the card tells whether or not the man has a tattoo on his face.

The elders' law may no longer be in effect; the best way to tell is to see whether any of the Kaluame men are breaking it. Indicate only those card(s) you definitely need to turn over **to see if any of these Kaluame men are breaking the law.**

The cards read: "eats cassava root", "no tattoo", "eats molo nuts", "tattoo".

(13) **Non-social contract permission rule—Fictitious culture:** Used in Experiments 6 and 9. For Experiment 9 the same text was used, but the rule was "switched".

You are an anthropologist studying the Namka, a hunter–gatherer culture living in the deserts of southwest Africa. The Namka people are divided into two great clans: the Bakas and the Heronas. Although the two clans are quite similar, it is easy to tell Bakas from Heronas due to a minor cultural quirk. As children, members of the Baka clan become adept at the secret art of finding ostrich eggshells, which they use as canteens because they are strong and hold lots of water. Herona children, however, become adept at the secret art of making canteens from goat skins; Herona children never learn how to find ostrich eggshells, just as Baka children never learn how to make goatskin canteens. Thus you can always tell the clans apart by seeing what kind of canteen a man has strapped to his side: a Baka carries an ostrich eggshell canteen, whereas a Herona carries a goatskin canteen. Members of the Baka

and Herona clans live together in peace and friendship. Important matters are decided by a group of Namka elders, half of whom are Baka, and the other half, Herona.

Twenty years ago, when you first started studying the Namka, the elders became concerned about the desert's dwindling animal population. The Namka hunt to get meat, and they particularly rely on two different species of small antelopes: duikers and gazelles. Duikers and gazelles are equally tasty and nutritious, and equally easy to hunt. Duikers are found only in the eastern part of the Namka's home range, whereas gazelles are found only in the western part. The problem is that both duikers and gazelles are in short supply, because the Namka population has been growing. If all the Namka lived and hunted in the eastern half of the home range, they would surely exhaust the supply of duikers; if they all lived and hunted in the western half, they would surely exhaust the supply of gazelles. The elders want their people to live in a balance with nature; they do not want to cause the extinction of either source of food, as this could eventually lead to the extinction of the Namka themselves.

Therefore, the elders decided to divide the Namka people in half, so that, roughly, one clan would live where the duikers roam, and eat only duiker meat, and one clan would live where the gazelles roam, and eat only gazelle meat. That way, neither food source would be overwhelmed by too many people, and both clans would be well nourished. Everyone agreed that this was a good plan. The only problem was that one clan had more people than the other, so 10% of the larger clan were asked to live, hunt, and eat with the smaller clan. They gladly agreed.

The elders expressed the law governing eating arrangements thus:

"If you eat duiker meat, then you have found an ostrich eggshell."

When you had left the Namka, everyone was happily observing this law, hoping that the duikers and gazelles would flourish as a result, so the next generation would not have to worry about such things.

That was 20 years ago. Now you are returning to southwest Africa to continue your study of the Namka, and you wonder whether the plan worked. Because the Namka are such law-abiding people, the best way to see if the law is still in effect is to watch what the Namka are eating; if any of them are breaking it, then it must be because the animals flourished and the elders repealed the law.

The cards below have information about four Namka men sitting in a temporary camp at the center of the home range; you can tell which clan each is from by their canteens. Each man is eating either duiker meat or gazelle meat, which he brought with him from home. Each card represents

one man. One side of a card tells which food a man is eating, and the other side of the card tells whether or not the man has ever found an ostrich eggshell.

The elders' law may no longer be in effect; the best way to tell is to see whether any of the Namka men are breaking it. Indicate only those card(s) you definitely need to turn over **to see if any of these Namka men are breaking the law.**

The cards read: "eats duiker meat", "has never found an ostrich eggshell", "eats gazelle meat", "has found an ostrich eggshell".

References

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

Axelrod, R., & Hamilton, W.D. (1981). The evolution of cooperation. *Science, 211*, 1390–1396.

Bracewell, R.J., & Hidi, S.E. (1974). The solution of an inferential problem as a function of the stimulus materials. *Quarterly Journal of Experimental Psychology, 26*, 480–488.

Brown, C., Keats, J.A., Keats, D.M., & Seggie, I. (1980). Reasoning about implication: A comparison of Malaysian and Australian subjects. *Journal of Cross-Cultural Psychology, 11*, 395–410.

Bruner, J.S. (1973). *Beyond the information given*, J.M. Anglin (Ed.). New York: Norton & Co.

Cheng, P., & Holyoak, K. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391–416.

Cheng, P., Holyoak, K., Nisbett, R., & Oliver, L. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18*, 293–328.

Chomsky, N. (1975). *Reflections on language*. New York: Random House.

Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.

Cosmides, L. (1985). *Deduction or Darwinian algorithms?: An explanation of the "elusive" content effect on the Wason selection task*. Doctoral dissertation, Harvard University. University Microfilms 86–02206.

Cosmides, L., & Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupre (Ed.), *The latest on the best: Essays on evolution and optimality*. Cambridge, MA: MIT Press.

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

Cox, J.R., & Griggs, R.A. (1982). The effects of experience on performance in Wason's selection task. *Memory and Cognition, 10*, 496–502.

Darwin, C. (1859/1958). *The origin of species*. New York: New American Library.

Dawkins, R. (1982). *The extended phenotype*. San Francisco: W.H. Freeman.

Evans, J.StB.T., & Lynch, J.S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*, 391–397.

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

Fodor, J.A. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

Gilhooly, K.J., & Falconer, W.A. (1974). Concrete and abstract terms and relations in testing a rule. *Quarterly Journal of Experimental Psychology, 26*, 355–359.

Golding, E. (1981, April). *The effect of past experience on problem solving*. Paper presented at the Annual Conference of the British Psychological Society, Surrey University.

Griggs, R.A., & Cox, J.R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73*, 407–420.

Griggs, R.A., & Cox, J.R. (1983). The effects of problem content and negation on Wason's selection task. *Quarterly Journal of Experimental Psychology, 35A*, 519–533.

Hamilton, W.D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology, 7*, 1–52.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69*, 366–378.

Holland, J.H., Holyoak, K.J., Nisbett, R.E., & Thagard, P.R. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.

Inhelder, B., & Piaget, J. (1958). *Growth of logical thinking: From childhood to adolescence*. New York: Basic Books.

Isaac, G.Ll. (1978). The food-sharing behavior of protohuman hominids. *Scientific American, 238*, 90–108.

Johnson-Laird, P.N. (1982). Thinking as a skill. *Quarterly Journal of Experimental Psychology, 34A*, 1–29.

Johnson-Laird, P.N., Legrenzi, P., & Legrenzi, M. (1972). Reasoning and a sense of reality. *British Journal of Psychology, 63*, 395–400.

Lee, R., & DeVore, I. (1968). *Man the hunter*. Chicago: Aldine.

Manktelow, K.I., & Evans, J.StB.T. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology, 70*, 477–488.

Manktelow, K.I., & Over, D.E. (1987). Reasoning and rationality. *Mind and Language, 2*, 199–219.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

Marr, D., & Nishihara, H.K. (1978). Visual information processing: Artificial intelligence and the sensorium of sight. *Technology Review*, October, 28–49.

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

Pollard, P. (1981). The effect of thematic content on the "Wason selection task". *Current Psychological Research, 1*, 21–29.

Pollard, P. (1982). Human reasoning: Some possible effects of availability. *Cognition, 10*, 65–96.

Reich, S.S., & Ruth, P. (1982). Wason's selection task: Verification, falsification and matching. *British Journal of Psychology, 73*, 395–405.

Rosenthal, R., & Rosnow, R.L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

Rozin, P. (1976). The evolution of intelligence and access to the cognitive unconscious. In J.M. Sprague & A.N. Epstein (Eds.), *Progress in psychobiology and physiological psychology*. New York: Academic Press.

Rozin, P., & Schull, J. (1988). The adaptive-evolutionary point of view in experimental psychology. In R.C. Atkinson, R.J. Herrnstein, G. Lindsey, & R.D. Luce (Eds.), *Handbook of Experimental Psychology*. New York: Wiley.

Rumelhart, D.E., & Norman, D.A. (1981). Analogical processes in learning. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.

Shepard, R.N. (1981). Psychophysical complementarity. In M. Kubovy & J.R. Pomerantz (Eds.), *Perceptual organization*. Hillsdale, NJ: Erlbaum.

Shepard, R.N. (1987). Evolution of a mesh between principles of the mind and regularities of the world. In J. Dupre (Ed.), *The latest on the best: Essays on evolution and optimality*. Cambridge, MA: MIT Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.

Staddon, J.E.R. (1987). Optimality theory and behavior. In J. Dupre (Ed.), *The latest on the best: Essays on evolution and optimality*. Cambridge, MA: MIT Press.

Stone, V. (1986). *The effect of causality on reasoning about conditional statements*. Unpublished manuscript, Stanford University.

Stone, V., & Cosmides, L. (in preparation). Do people have causal reasoning schemas?

Symons, D. (1987). If we're all Darwinians, what's the fuss about? In C. Crawford, D. Krebs, & M. Smith (Eds.), *Sociobiology and psychology*. Hillsdale, NJ: Erlbaum.

Tooby, J. (1985). The emergence of evolutionary psychology. In D. Pines (Ed.), *Emerging syntheses in science*. Santa Fe: Sante Fe Institute.

Tooby, J., & DeVore, I. (1987). The reconstruction of hominid behavioral evolution through strategic modeling. In W.G. Kinzey (Ed.), *The evolution of human behavior: Primate models.* New York: SUNY Press.

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

Trivers, R.L. (1972). Parental investment and sexual selection. In B. Campbell (Ed.), *Sexual selection and the descent of man, 1871–1971.* Chicago: Aldine.

Trivers, R.L. (1974). Parent–offspring conflict. *American Zoologist, 14*, 249–264.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232.

Van Duyne, P.C. (1974). Realism and linguistic complexity in reasoning. *British Journal of Psychology, 65*, 59–67.

Van Duyne, P.C. (1976). Necessity and contingency in reasoning. *Acta Psychologica, 40*, 85–101.

Wason, P.C. (1966). Reasoning. In B.M. Foss (Ed.), *New horizons in psychology.* Harmondsworth: Penguin.

Wason, P.C. (1983). Realism and rationality in the selection task. In J.StB.T. Evans (Ed.), *Thinking and reasoning: Psychological approaches.* London: Routledge & Kegan Paul.

Wason, P.C., & Johnson-Laird, P.N. (1972). *Psychology of reasoning: Structure and content.* London: Batsford.

Wason P.C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology, 23*, 63–71.

Williams, G.C. (1966). *Adaptation and natural selection: A critique of some current evolutionary thought.* Princeton, NJ: Princeton University Press.

Yachanin, S.A., & Tweney, R.D. (1982). The effect of thematic content on cognitive strategies in the four-card selection task. *Bulletin of the Psychonomic Society, 19*, 87–90.

Résumé

Pour s'engager avec succès dans un échange social—une coopération avec bénéfice mutuel entre deux ou plusieurs individus—les êtres humains doivent pouvoir résoudre avec efficacité un certain nombre de problèmes computationnels complexes. Marr (1982), Cosmides (1985) et Cosmides et Tooby (1988) ont utilisé des principes évolutionnistes pour développer une théorie computationnelle de ces problèmes d'adaptation. Des hypothèses spécifiques sur la structure des algorithmes qui gouvernent la façon dont les humains raisonnent à propos des échanges sociaux dérivent de cette théorie. Dans cet article on présente une série d'expériences destinées à tester ces hypothèses en utilisant un test de raisonnement logique: la tâche de sélection de Wason. Dans la première partie on présente les expériences testant la validité de la théorie de l'échange social (social exchange theory) contre celle des théories de raisonnement; dans la deuxième partie on rapporte les expériences testant cette théorie contre la théorie du schéma de permission (permission schema theory) de Cheng et Holyoak (1985). Le plan expérimental inclut huit tests critiques concus pour départager la théorie de l'échange social des deux autres familles de théories; les résultats des huit tests sont tous en faveur de la théorie de l'échange social. Les effets de contenu trouvés dans ces expériences appuient l'hypothèse que l'esprit humain inclut des processus cognitifs spécialisés pour raisonner dans l'échange social et n'appuient que parcimonieusement les processus rapportés dans la littérature. On discute les implications de cette ligne de recherche pour une vue modulaire de l'esprit humain et pour l'utilité de la biologie évolutionniste dans le développement des théories computationelles.