2 Can a General Deontic Logic Capture the Facts of Human Moral Reasoning? How the Mind Interprets Social Exchange Rules and Detects Cheaters

Leda Cosmides and John Tooby

I Evolutionary Psychology and Deontic Logic

The recognition that our cognitive and motivational architecture is the product of natural selection raises the possibility that our moral concepts, moral intuitions, and moral sentiments might themselves be reflections of the evolutionary process. Indeed, this conclusion seems difficult to escape, given how natural selection works.

Using what is known about the evolutionary process and the behavioral ecology of ancestral hunter-gatherers, we initiated a research program in the early 1980s aimed at discovering whether the human mind reliably develops an evolved specialization for deontic reasoning about social exchange. Social exchange—also called "reciprocity," "reciprocal altruism," or "trade," depending on the research community—is cooperation for mutual benefit between two agents. Starting with evolutionary game theory, where it is analyzed as a repeated Prisoners' Dilemma, we developed *social contract theory*: a task analysis of the computational requirements for adaptively engaging in social exchange (see Cosmides, 1985; Cosmides & Tooby, 1989). Many of these requirements were so particular to adaptive problems that arise in social exchange that they could only be implemented by a computational system whose design was functionally specialized for this function. To discover whether a system of this kind exists in the human mind, we conducted reasoning experiments that looked for evidence of the design features predicted by social contract theory.

Our question was, does the human cognitive architecture reliably develop *social contract algorithms*: a neurocomputational system whose design features are adaptively specialized for producing the specific kinds of inferences and goals necessary to create cooperative interactions that implement an evolutionarily stable strategy (ESS)? By hypothesis:

1. Social contract algorithms are activated by particular cues indicating a criss-crossing pattern of desires, access to benefits, and intention to provide benefits that characterizes situations involving social exchange.
2. Once activated, social contract algorithms represent situations via a proprietary format that represents distinctions that were adaptively important in the domain for which they evolved (e.g., *agent*$_i$ *benefit to agent*$_i$, *requirement of agent*$_j$, *obligations*, *entitlements, cheating*).
3. Social contract algorithms are equipped with functionally specialized inferential procedures that were designed to operate on these proprietary representations, generating inferences that, while not true across domains, were adaptively useful when operating within the system's proper domain of application. Some of these inferences apply rules of transformation specific to social exchange, which allow obligations to be inferred from entitlements and vice versa—a form of deontic reasoning. Others regulate the detection of cheaters. The ability to detect cheaters is necessary to implement an ESS for social exchange. Thus, the adaptive function of a cheater detection mechanism would be to search for information that could reveal who has committed a specific kind of moral violation: intentional cheating in a situation of social exchange.

We have spent almost 25 years empirically testing the predictions of social contract theory and will review some of the evidence for it. In doing so, we will focus on results that distinguish reasoning about social exchange from reasoning about conditional rules drawn from other deontic domains, especially precautionary rules. We will do this for two reasons. First, casual readers of the reasoning literature have developed the mistaken belief that results found for reasoning problems involving social exchange are found for all deontic rules. This is untrue. Second, differences in how people reason about deontic rules drawn from different domains may have implications for a project within moral philosophy: the development of a domain-general deontic logic.

I.i Why Should Moral Philosophers Care about Human Nature?

I.i.i Natural Selection Is an Amoral Process, yet It Can Produce Moral Intuitions Natural selection favors designs on the basis of how well they promote their own reproduction, not on how well they promote moral behavior. If this is not obvious, consider the fate of a mutation that alters the development of a neural circuit, changing its design away from the species standard. This new circuit design implements a decision rule that produces a radically different moral choice in a particular type of situation: help rather than hurt, cooperate rather than cheat. Will this new decision rule, initially present in one or a few individuals, be eliminated from the population? Or will it be retained, increasing in frequency over the generations until it replaces the old design, eventually becoming the new "standard model" in that species?

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

This is the process that, over eons, constructed human nature—that is, the reliably developing, species-typical information-processing architecture of the human mind. As a result, human nature is comprised of programs that were selected for merely because they outreproduced alternative programs in the past. There is nothing in this process to ensure the production of decision rules or moral sentiments that track the desiderata of an ethically justifiable moral system. So why should moral philosophers care about human nature?

I.i.ii Human Nature, Evolved Inferences, and Moral Philosophy Human nature is relevant to moral philosophy for many reasons (Cosmides & Tooby, 2004, 2006), but here we will be concerned with just two of them. Moral philosophers should care about human nature first and foremost because they themselves are members of the human species. If the human cognitive architecture contains programs that generate moral intuitions and inferences in humans, then it generates moral intuitions and inferences in humans who are moral philosophers. These evolved programs cause specific moral inferences to be triggered by particular situations, whether those situations are occurring in real life or merely in imagination (Boyer, 2001; Cosmides & Tooby, 2000a). Counterfactual and suppositional

W

arguments are important tools of the moral philosopher, but counterfactual propositions are held in metarepresentations to which evolved, domain-specific inference procedures are then applied (Leslie, 1987; Cosmides & Tooby, 2000b). Counterfactuals and their downstream inferences are thereby decoupled from the encyclopedia of world knowledge stored in one's semantic memory, preventing the kind of data corruption that Leslie (1987) has called "representational abuse" (Cosmides & Tooby, 2000b). This means that deliberative reasoning is not immune from the influence of domain-specific evolved programs (see, e.g., Lieberman, Tooby, & Cosmides, 2003; Haidt, 2001, on moral dumbfounding; Gendler, 2000, 2003, on affective transmission and imaginative resistance; Tooby & Cosmides, 2001).

As a result, evolved social inferences and moral intuitions can be expected to affect the inferences and judgments that moral philosophers make, even when they are making counterfactual or suppositional arguments that engage more domain-general metarepresentational machinery. Sometimes moral sentiments in response to counterfactual arguments will be nothing more than readouts of evolved programs that were generated by an amoral process—hardly a ringing endorsement. Moral philosophers need to recognize when their judgments are based on evolved moral sentiments and inferences and need to decide how this should factor into their theories.

A more subtle problem arises when one realizes that certain concepts themselves are products of the evolutionary process and were selected for because of the way they interacted with motivational systems (Tooby, Cosmides, & Barrett, 2005). This is true of certain moral concepts, such as the deontic notions of *obligation* and *entitlement*. A single word, such as "ought," "should," "obligated," or "must," may map onto several different evolved concepts, each embedded in a different, domain-specific inferential system. "Must," for example, refers to a different underlying concept when it appears in the context of social exchange than when it appears in a precautionary rule, as we will discuss below. This has implications for certain projects in logic and moral philosophy—which brings us to the second reason philosophers should care about the evolved architecture of the human mind.

For more than half a century, philosophers have been trying to develop a deontic logic that satisfies two goals. The first goal is to capture "the logical structure of our ordinary deontic language and . . . our ordinary deontic reasonings" (Castañeda, 1981, p. 38). The second is to create a formal calculus, a syntax that applies deontic concepts such as "obligation" in a content- and context-independent manner[1] (von Wright, 1951;

Hilpinen, 1971; McNamara, 2006). Feldman (2001, p. 1011), for example, expresses the latter hope when he says that "the plausibility of any logical claim is enhanced if it can be seen to cohere with an overarching conception of the logic of obligation." In suggesting a method for evaluating candidate deontic operators, Feldman says that in deontic logic systems

> . . . some operator (usually "O") is intended to be the formal analog of "ought" in one of its ordinary language senses. *The systematic logical features of the operator are precisely determined*. We then consider the extent to which the logical features of the formal operator correspond to those we intuitively suppose belong to "ought" in ordinary discourse. (Feldman 2001, emphasis added)

This passage acknowledges that "ought" has multiple senses in ordinary language. Yet it expresses the hope that one can determine *one set of systematic logical features* of operator "O" while still finding that these features correspond to people's ordinary intuitions about what "ought" implies. Accomplishing this should be difficult if "ought" has many senses in ordinary language: it would entail either choosing one sense and ignoring the others or discovering that what at first appear to be a multiplicity of meanings and implications really collapse onto one set that applies across contexts.

The first goal—capturing ordinary reasoning about deontic concepts—implies that the construction of a deontic logic must be constrained by empirical data on language use and deontic reasoning. Psychological data are clearly relevant to that goal, so we will review our research on deontic reasoning. But that research suggests that deontic reasoning is not a unified phenomenon. If so, then the first goal may be incompatible with the second goal: constructing a deontic logic with operators such as "O" that apply across all human contexts.

Creating a domain-general deontic logic would require deontic operators and rules of inference that apply in a uniform way across every (deontic) context involving human action. This includes social, moral, legal, religious, and prudential contexts (at minimum). But what if a single lexical item, such as "ought," "obligated," or "entitled," masks a plethora of meanings that bear only a family resemblance to one another? For example, what if the *ought* embedded in social contract algorithms has a different meaning/set of implications than the *ought* embedded in a precautionary inference system—that is, what if these concepts are better thought of as $\textit{ought}_{SC}$ and $\textit{ought}_{Prec}$?

We suspect this is a real possibility, for two reasons. The first is empirical: deontic reasoning seems to fractionate into functionally distinct domains, as we will discuss below. The second reason is theoretical and involves how natural selection tends to engineer evolved systems.

I.i.iii Domain Specificity, Evolution, and Deontic Reasoning What counts as adaptive social behavior differs by domain. Courtship, dyadic exchange, n-person cooperation, deep engagement friendship, dominance relationships, coalitional versus individual aggression, parent-child relationships, sibling relationships, hazard management—each is associated with a different set of adaptive information-processing problems, some of which are unique to that domain (e.g., Bugental, 2000; Buss, 1994; Cosmides & Tooby, 1987, 1989; Fiddick, Cosmides, & Tooby, 2000; Fiske, 1991; Kurzban, McCabe, Smith, & Wilson, 2001; Symons, 1987; Tooby & Cosmides, 1996; Tooby, Cosmides, & Price, 2006; Trivers, 1974). In saying that two adaptive problems are different, we mean something specific: that a neurocomputational system whose design is well engineered for solving one will not be well engineered for solving another (for extended discussion, see Tooby & Cosmides, 1992). When adaptive problems differ by domain, natural selection tends to produce different neurocomputational systems as solutions to these problems, each equipped with domain-specialized design features (for some stunning examples, see Gallistel, 2000). This perspective suggests that social interaction in humans will be regulated by a number of different evolved specializations, each of which is functionally specialized for negotiating a particular domain of social life.

Social interaction across many domains involves moves and countermoves, expectations, obligations, prohibitions, and entitlements. But the nature of these can be very different depending on whether one is interacting with a mate, a sibling, a child, an exchange partner, a dear friend, a status rival, a chief/superior, a foe, or a comrade-at-arms. This means that several evolved specializations, each domain specialized and context specific, may employ some *version* of a particular deontic concept. But each version may differ from the others by virtue of the unique inferential role it plays within its particular evolved inference system.

If this picture is even remotely correct, then the project of creating a deontic logic that is both general yet empirically descriptive may be doomed. Deontic logicians insisting on domain generality would be driven to define deontic concepts in a manner so general as to be useless—a manner that does not escape the problems of domain specificity but merely hides them.

Consider, for example, the very general definition of "ought" as "A person P ought to take action A when P has a reason to take A" (or "when there is a reason to take A"; e.g., Mackie, 1977). This appears general, but it is not: all the heavy lifting is shifted onto another word, in this case, onto what, exactly, it means "to have a reason to." "Having a reason to" could refer to having a mental representation of there being (i) means sufficient to attain a physical goal, (ii) means sufficient to attain a social goal (whether moral or not; e.g., helping a neighbor, having a good marriage, attaining dominance over others), (iii) means *believed* to help attain a goal (whether efficacious or not; e.g., sacrificing a goat to appease the gods), (iv) an ethical obligation, (v) prudential advice about how to reduce risk. "Having a reason to" could also refer to (vi) the wish to be (or appear) pious, (vii) the wish to conform to the requirements of a legal system, (viii) the wish to stay out of prison, (ix) the wish to avoid ostracism or social opprobrium, (x) the (somewhat different) wish to be seen as a good community member, and so on. Moreover, whatever its content, the mental representation of that "reason" could be in a form that allows conscious awareness, reflection, and verbal report or it could be implicit in the procedures and logic of an evolved program. In either case, the domain-specificity of "ought" creeps back in, under different cover.

A more realistic—and more illuminating—philosophical goal may be to embrace the domain and context specificity of moral concepts. The meaning of these concepts could be worked out, taking into account the role each plays in an evolved inferential system, as well as other more philosophical desiderata, such as noncontradiction, consistency, and so on. The project would not be to develop a domain-general deontic logic on analogy to the domain-general alethic logics. The project would instead be to develop a series of very well-specified domain-specific deontic logics, each of which applies within certain boundaries of human action and not outside them.[2]

Truth-preserving logics are not always good descriptions of how people intuitively reason. But who cares? They are still useful for increasing knowledge, whether they are implemented by a computer system or by a human being who is reasoning deliberatively and laboriously, using pencil and paper to store intermediate inferences and conclusions. Deontic logics can also be useful, but for a different reason: they can clarify the moral dimension of human affairs. Deontic logicians are certainly aware of this:

> . . . despite the fact that we need to be cautious about making too easy a link between deontic logic and practicality, many of the notions listed are typically employed in attempting to regulate and coordinate our lives together (but also to evaluate states of affairs). For these reasons, deontic logics often directly involve topics of considerable practical significance such as morality, law, social and business organizations

> (their norms, as well as their normative constitution), and security systems. (McNamara, 2006)

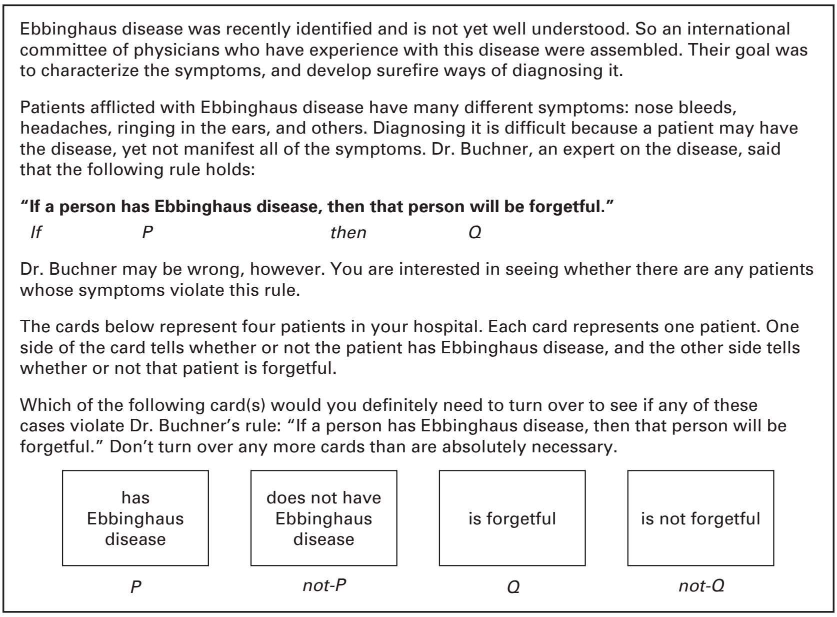
The prospect of improving human affairs is a heady possibility. But how useful will a deontic logic be in this regard if it fails to capture major distinctions the human mind makes when reasoning deontically?

The fact that natural selection shaped certain mechanisms for moral reasoning does not justify them, but a formal deontic calculus that deeply violates our moral intuitions is not likely to be widely understood or adopted (Boyer, 2001; Sperber, 1996). Without being widely understood and adopted, a deontic logic will not succeed in guiding ethical decisions beyond an esoteric circle of specialists.

Deontic logics have the potential to illuminate and clarify moral reasoning, but to have a real impact on human affairs, they need to satisfy both normative and descriptive goals. With these thoughts in mind, we will review what has been learned about the deontic logic that our minds deploy in situations of social exchange. To situate these findings in the larger intellectual landscape, we begin with a brief overview of how psychologists have approached the study of reasoning, focusing on a tool used extensively in our investigations, the Wason selection task.

I.ii Traditional Conceptions of Rationality in the Study of Human Reasoning

With the cognitive revolution, psychologists began to reverse engineer the mechanisms by which the human mind reasons. The goal was to discover what representations and inferential rules these programs apply. When this enterprise began, traditional views of rationality dominated psychological research. The first hypothesis considered was that programs that cause human reasoning implement "rational algorithms": ones that embody normative, truth-preserving rules of inference derived from formal logic (especially first-order logic) or mathematics (e.g., Bayes's theorem). For example, Peter Wason, a pioneer in the study of the psychology of reasoning, explored the notion that everyday learning was a form of Popperian hypothesis testing: one projects a hypothesis, framed as a conditional rule, and then seeks falsifying instances. His four-card selection task was designed to see whether people would spontaneously and accurately look for potential violations of a conditional rule, linguistically expressed as *If P then Q* (see figure 2.1). He found, much to his surprise, that they did not. By first-order logic, *If P then Q* is violated by any instance or situation in which *P* is true and *Q* is false, so the solution to Wason's selection



Ebbinghaus disease was recently identified and is not yet well understood. So an international committee of physicians who have experience with this disease were assembled. Their goal was to characterize the symptoms, and develop surefire ways of diagnosing it.

Patients afflicted with Ebbinghaus disease have many different symptoms: nose bleeds, headaches, ringing in the ears, and others. Diagnosing it is difficult because a patient may have the disease, yet not manifest all of the symptoms. Dr. Buchner, an expert on the disease, said that the following rule holds:

"If a person has Ebbinghaus disease, then that person will be forgetful."
If *P* *then* *Q*

Dr. Buchner may be wrong, however. You are interested in seeing whether there are any patients whose symptoms violate this rule.

The cards below represent four patients in your hospital. Each card represents one patient. One side of the card tells whether or not the patient has Ebbinghaus disease, and the other side tells whether or not that patient is forgetful.

Which of the following card(s) would you definitely need to turn over to see if any of these cases violate Dr. Buchner's rule: "If a person has Ebbinghaus disease, then that person will be forgetful." Don't turn over any more cards than are absolutely necessary.

has Ebbinghaus disease	does not have Ebbinghaus disease	is forgetful	is not forgetful
P	*not-P*	*Q*	*not-Q*

Figure 2.1
The Wason selection task (indicative, descriptive rule, familiar content). In a Wason task, there is always a rule of the form *If P then Q*, and four cards showing the values *P*, *not-P*, Q, and *not-Q* (respectively) on the side that the subject can see. By first-order logic, only the combination of *P* and *not-Q* can violate this rule, so the correct answer is to check the *P* card (to see if it has a *not-Q* on the back), the *not-Q* card (to see if it has a *P* on the back), and no others. Few subjects answer correctly, however, when the conditional rule is descriptive (indicative), even when its content is familiar; for example, only 26% of subjects answered the above problem correctly (by choosing "has Ebbinghaus disease" and "is not forgetful"). Most choose either *P* alone or *P* & Q. (The italicized *P*s and *Q*s are not in problems given to subjects.)

task is to choose the *P* card (to see if it says *not-Q* on the back) and to choose the *not-Q* card (to see if it has a *P* on the back). But when the conditional rule was indicative, purporting to describe some relationship in the world, only 5%–30% of people chose *P*, *not-Q*, and no other cards. Most chose *P* alone or *P* & *Q*, behavior that is consistent with a confirmation bias. Taking a semester-long course in logic did not improve students' performance (Cheng, Holyoak, Nisbett, & Oliver, 1986). Wason even found that logicians sometimes got it wrong at first, admitting in retrospect that they should have chosen *P* & *not-Q* (Wason & Johnson-Laird, 1972).

Follow-up studies eliminated many possible reasons for this poor performance. It's not that people are interpreting the rule as a biconditional and then reasoning logically: that would lead them to choose all four cards, which is a rare response. It's not that people require more than one violating instance to decide the rule is false: the same levels of performance prevail when subjects are only asked to look for potential violations of the rule, without being asked to evaluate the truth of the rule. It's not that the indicatives tested involved an arbitrary relation between letters and numbers. During the late 1970s and early 1980s, indicatives involving familiar, everyday content were tested. Wason performance on indicative conditionals remained low whether the terms were concrete and familiar, the relation between them was concrete and familiar, or both (e.g., Cosmides, 1985, 1989; Griggs & Cox, 1982; Manktelow & Evans, 1979; Wason, 1983; Yachanin & Tweney, 1982; reviewed in Cosmides, 1985). This should be obvious from the conditional rule shown in figure 2.1. What could be more familiar or "natural" than the notion that a disease causes symptoms? Yet this conditional rule elicited the correct response from only 26% of undergraduates. (Buller's, 2005, recent claim that people do well on the Wason task when the terms or relations of an indicative are familiar or "natural" has been known to be false for many years.)

Most interestingly, poor performance on the selection task is not because people don't know that an instance of *P & not-Q* would violate the rule. When people are asked whether a case of *P & not-Q* would violate the rule (an evaluation task), they recognize that it would (Manktelow & Over, 1987). To investigate in another way the relation between knowing what counts as a violation and searching for one, we once did a series of Wason studies in which we first told subjects exactly what counts as a violation. In these studies, subjects were given a conditional rule purporting (e.g.) to describe people's transportation habits: *If a person goes into Boston, then that person takes the subway*. The cards represented four people, and each specified where that person went on a particular day and how they got there. The facing side of the cards read *Boston*, *Arlington*, *subway*, and *cab*. Subjects were asked to indicate those card(s) they would need to turn over to see if the behavior of any of these individuals violates the rule—so far, a standard selection task. But at the top of the page, we first explained—in ever increasing detail—what would count as violating a conditional rule of this kind. In one condition we put it abstractly, explaining that a rule of the form *If X happens, then Y happens* "has been broken whenever *X* happens but *Y* does not happen. Remember this in answering the question below." In another condition we added to this: "So, for example, the rule below would be broken by a person who went to Boston by cab." In yet another version, we added: "So, for example, the rule below would be broken by a person who went to Boston without taking the subway—e.g., a person who went to Boston by cab." (Note that these last two mention the precise cases of *P* and *not-Q* shown on the cards.) In all cases, performance hovered around 25% correct—no different from a condition in which we said nothing about what counts as a violation. (We did a parallel series using a deontic rule that was not a social contract; results were the same, with performance averaging 18% correct.) Combined with results from evaluation tasks, this is telling: although people *recognize* what counts as violating an indicative rule, they do not *seek out* information that could tell them whether the rule has been violated, even when they are explicitly asked to do so and have all the time they wish (see also Fiddick et al., 2000, pp. 44–45, 74–75).

Why was poor performance on Wason's selection task significant? Even the simplest computer programming languages have inferential rules that operate on the syntax of If-then, of antecedents and consequents, regardless of their content. Basic rules of logic that operate on this syntax, like *modus ponens* and *modus tollens*, are useful precisely because they allow true conclusions to be derived from true premises, no matter what the content of the premises may be. That the human brain might be equipped with such rules, operating over a content-general If-then syntax, is not only reasonable, but it fits with prevailing notions that we are the "rational animal," the species where reason emerged, erasing instincts in its wake. If the human brain, like the simplest of computer languages, is equipped with logical rules of inference operating on the syntax of If-then, then why weren't more people succeeding on the Wason selection task? *Modus ponens* would lead them to choose the *P* card, *modus tollens* to choose the *not-Q* card. Moreover, the problem space for the Wason task is finite and small: regardless of which face is up, the possible values on the cards are *P & Q*, *P & not-Q*, *not-P & Q*, *not-P & not-Q*. Given that people recognize that *P & not-Q* counts as a violating instance, an exhaustive search of this small problem space should lead them to choose the only cards that could combine these two values: the *P* card and the *not-Q* card. Indeed, they do choose the *P* card, almost always, consistent with other data suggesting that the brain's neural circuitry does implement and spontaneously apply a *modus ponens* inference rule to the syntax of If-then (e.g., when one is asked to draw a conclusion from premises or to evaluate the validity of a conclusion drawn by others; Rips, 1994; Wason & Johnson-Laird, 1972). But the *not-Q* card is chosen only about 50% of the time, consistent with other research showing that *modus tollens* is not spontaneously applied as often as *modus ponens* is (Rips, 1994; Wason & Johnson-Laird, 1972). In addition, cards inconsistent with either logical rule of inference—especially the *Q* card—are often chosen. Why does this happen? The task is conceptually so simple, the solution could be implemented on a computer in a few lines of code.

Selection task results for indicative conditionals led many to doubt that the brain has circuitry that implements all the rules of first-order logic. That debate remains unsettled: perhaps it does, but only within a comprehension module (Sperber, Cara, & Girotto, 1995); perhaps pre-existing beliefs hijack logical analysis (Evans, 1989); perhaps the search for confirming instances is a more useful learning strategy when someone proposes *If P then Q*, given a world where there are an infinite number of properties that have no association with the presence of property *P*, and only a few properties that do.

I.ii.i Deontic Exceptions to the Pattern Poor performance on the Wason task was first noted with indicative rules that relate numbers and letters, and then with indicatives that link more familiar terms and relations, such as "If a person eats hot chili peppers, then he will drink a cold beer"—despite the familiar experience of trying to sooth a burning mouth with a cold drink, this particular rule elicited only 35% correct from Harvard undergraduates, even when *not-Q* referred to drinking hot tea (Cosmides, 1985). The failure to answer "*P & not-Q*" was so pervasive that it was initially thought to apply to all conditional rules. But during the early to mid 1980s, it became apparent that one could elicit "*P & not-Q*" responses with tasks employing a deontic conditional, that is, a conditional statement describing what a person is *obligated* or *entitled* to do in a given context. An early example was from Griggs and Cox (1982), who elicited high levels of "*P & not-Q*" responses with the drinking age problem ("If a person is drinking beer, then the person must be over 19 years old" (*N.B.* 19 was the legal drinking age at the time.)). But high performance with this rule was not initially attributed to the fact that it is deontic rather than indicative. This law was so well-known (and so famously violated) by undergraduate subjects that excellent performance was thought to reflect the familiarity of the rule and/or the fact that many subjects had personal experiences of having violated it, which they could retrieve from memory and match to the cards. That good violation detection on the Wason task could be elicited *systematically* by certain deontic rules first became apparent from our own work (Cosmides, 1985, 1989; Cosmides & Tooby, 1989) and that of Cheng and Holyoak (1985; Cheng, Holyoak, Nisbett, & Oliver, 1986). The theoretical perspectives were quite different, but the phenomena similar: *certain* deontic rules could elicit good violation on the selection task, regardless of their familiarity.

One set of results was presented in Cosmides's (1985) dissertation and then circulated like samizdat for a number of years, eliciting published commentary (e.g., Manktelow & Over, 1987) several years before appearing in journals. (The fact that the hypotheses tested were derived from theories originating in evolutionary biology was considered inflammatory at the time and was a deal breaker for many psychology journals.) We had proposed that the human mind reliably develops *social contract algorithms*: computational machinery that is functionally specialized for reasoning about social exchange (Cosmides, 1985, 1989; Cosmides & Tooby, 1989). In this work, we showed that deontic rules that fit the template of a social contract (see below) reliably elicit high levels of violation detection on the Wason task, that they do so regardless of the rule's familiarity, that the rules of inference are specific to social exchange and do not map onto rules of inference of first-order logic, and that deontic rules that do not fit the social contract template do not elicit high levels of violation detection.[3] Because social exchange involves deontic concepts, social contract algorithms embody a form of deontic logic (Sections II and III). But their logic and operations are not general to all deontic rules: They are well engineered only for interpreting and reasoning about situations involving social exchange, and their ability to detect violations is restricted to those produced by intentional cheating (Section IV).

At the end of 1985, Cheng and Holyoak published an influential paper also showing that certain deontic rules can systematically elicit high levels of violation detection. The scope of their explanation—permission schema theory—is broader than the scope of social contract theory. If it were true, the prospects for creating a deontic logic that is content general yet still respects empirical data about the fracture points of human reasoning would be more promising. In Section IV, we will discuss data that speak against their view and others like it, showing that "deontic reasoning" fractionates into a number of different domains.

I.iii Ecological Rationality in the Study of Reasoning

The traditional view held human reasoning to be rational to the extent that it conforms to normative theories drawn from mathematics and logic. Our research on reasoning emerged from a different view, sometimes called *ecological rationality*. According to this view, the human cognitive architecture reliably develops a number of functionally isolable computational systems, each of which is specialized for solving a different adaptive problem (for reviews, see Barkow, Cosmides, & Tooby, 1992; Boyer, 2001; Buss, 2005; Hirschfeld & Gelman, 1994; Pinker, 1997; Gallistel, 2000). When activated by content from the appropriate domain, these inference engines impose special and privileged representations during the process of situation interpretation, define specialized goals for reasoning tailored to their domain, and make available specialized inferential procedures that allow certain computations to proceed automatically or "intuitively" and with enhanced efficiency over what a more general reasoning process could achieve given the same input (Cosmides & Tooby 1987, 2005a; Tooby & Cosmides, 1992, 2005). The designs of these systems may not embody content-independent norms of rationality, such as the predicate calculus or Bayes's rule, but they are *ecologically rational* (Gigerenzer, Todd, & ABC Research Group, 1999; Cosmides & Tooby, 1996a; Tooby & Cosmides, in press). That is, each embodies functionally specialized design features that reflect the task demands of the adaptive problem it evolved to solve, including assumptions about the evolutionarily long-term ecological structure of the world. As a result, when operating within the domain for which they evolved, these systems solve adaptive problems reliably, efficiently, and with limited information.

The Charlie task, from Simon Baron-Cohen's (1995) research on theory of mind, provides a good illustration of an ecologically rational inference system. It reveals the presence of a reasoning mechanism that is useful for inferring people's goals and desires because its structure reflects an evolutionarily long-enduring feature of the world: that people, like other animals, often turn their eyes in the direction of objects they are interested in.

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

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

In what sense is this reasoning instinct domain specific, content specific, and ecologically rational? It is domain specialized because it is well engineered for making inferences about agents and their mental states, not about plants, mountains, or tools—not even about aspects of agents unrelated to their mental states (not their substance, their appropriateness as projectiles, etc.). It is content specialized for several reasons. (1) It takes information about eye direction and object-of-eye-direction as input. (2) It has specialized representational formats: it represents people as having (invisible) mental states like *wanting* that cause their behavior and infers the content of these mental states (*wants*(Milky Way)). And (3) the inferential procedure activated has nothing to do with logic or mathematics: it is specific to eye gaze and mental states. What makes it ecologically rational? When operating within its proper domain (agents and mental states), it produces inferences that are likely to be true because they are based on an ecological regularity of past[4] environments: that eye direction predicts object of attention. When operating outside its proper domain (e.g., when activated by "eyes" on a butterfly's wings), the inferences it produces are useless.

To create a content-free inference system, one whose procedures operate validly across all domains, one would need to *remove* this very useful reasoning instinct precisely because it produces useful inferences in some domains but not in others. As a result, an inference system that is completely content free would be computationally weaker than a system that is replete with ecologically rational inference procedures like this one: content-general[5] and content-specific ones, all working together.

Ecologically rational inference systems like this one gain their inferential power by taking advantage of relationships that are true (or at least statistically reliable) within a particular problem domain, and they solve specific problems that arise within that domain. These problems need not be about determining what is true of the world; they may be inferences about how to behave. A snake avoidance system isn't just about whether a snake is present; it includes the inference that one should move away from the snake (see Tooby, Cosmides, & Barrett, 2005, on how selection pressures shaping motivational systems should have shaped co-adapted conceptual systems). To solve these domain-limited problems, each of these inference systems should be designed to operate over a different and limited class of content and be activated by cues associated with their proper domain (Barrett, 2005; Sperber, 1994). On this view, we should expect to find that the human cognitive architecture is densely populated with a large number of evolved, content-specific, domain-specific inference engines (or evolved mechanisms for their developmental acquisition), in addition to whatever more domain- or content-general inferential competences may exist.

We will be discussing two computational systems of this kind. One is designed for reasoning about social exchange: for detecting when a situation involves social exchange, for making appropriate inferences about what these interactions entail, and for detecting cheaters (Cosmides, 1985, 1989; Cosmides & Tooby, 1989, 1992, 2005a). The other is designed for reasoning about risk reduction in hazardous situations and is well engineered for detecting violations of precautionary rules—instances in which a person might be in danger by virtue of having failed to take appropriate precautions (Boyer & Lienard, in press; Cosmides & Tooby, 1997; Fiddick, 1998; Fiddick et al., 2000; Leckman & Mayes, 1998, 1999; Stone, Cosmides, Tooby, Kroll, & Knight, 2002). These systems make certain inferences easy, effortless, and intuitive when one is reasoning about problems that tap their respective domains.

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

This problem—we call it "instinct blindness"—can afflict anyone, including philosophers. Occasionally, we encounter philosophers who try to account for the data we discuss below by saying that social contracts (along with other deontic rules) "have" a different interpretation than indicative rules (Buller, 2005; Fodor, 2000). This is naive realism. Social contracts do not "have" an interpretation; the mind *assigns* an interpretation to them, and it does so when certain cues and conditions are contextually present (Cosmides, 1985, 1989; Cosmides & Tooby, 1992; Fiddick et al., 2000; Gigerenzer & Hug, 1992; Platt & Griggs, 1993). Social contract theory is an account of the computational procedures *by which these interpretations are assigned*: what cues are necessary and/or sufficient, which inferential transformations these procedures license, and so on. Social contract theory predicted—in advance of the empirical evidence—that people will interpret conditionals expressing social exchange differently from other conditionals (both indicative and deontic) and make different inferences about them. That prediction, now empirically validated, is one of the facts that the theory predicted and explains. Social contract theory is an attempt to replace the black box of "interpretation" with computational machinery that does the interpreting—in short, to build a cognitive science of central processes.

We hope philosophers will spill no more ink explaining to us that interpretion plays a role in reasoning about conditional rules, including social contracts. Explaining how the mind interprets conditional rules has been an important aspect of social contract theory since its inception (Cosmides, 1985; Cosmides & Tooby, 1989). Indeed, a 2000 paper of ours highlights this in the title: "No Interpretation without Representation: The Role of Domain-Specific Representations and Inferences in the Wason Selection Task" (Fiddick et al., 2000).[6] A more productive use of time would be to propose a theory of interpretation that differs from social contract theory's, yet explains reasoning performance. In attempting to do so, however, two facts must be kept in mind, which we will elaborate on below. First, many deontic rules *do not elicit good violation detection on the Wason task*. This means that no explanation that refers to interpretation can be correct if it is general to all deontic rules. Second, whether a given social contract rule elicits good violation detection depends on whether the context proposes that the violation results from intentional cheating or an innocent mistake. *Yet the interpretation of the rule is the same in both cases*: it is precisely the same rule embedded in a story that gives each of its terms and the relation between them precisely the same social contract meaning. All that differs are the proposed motivations of the potential violator. This result is important: it means that assigning a "social contract" interpretation to a conditional rule is not sufficient for eliciting good violation detection.

That said, we turn to the theory.

II Evolutionarily Stable Strategies and Social Contract Theory

Social contract theory is based on the hypothesis that the human mind was designed by evolution to reliably develop a cognitive adaptation specialized for reasoning about social exchange. To test whether a system is an adaptation that evolved for a particular function, one must produce design evidence. It is an engineering standard: functional design is evidenced by a set of features of the phenotype that (i) combine to solve an element of a specific adaptive problem particularly well and (ii) do so in a way unlikely to have arisen by chance alone or as a side effect of a mechanism with a different function. Hence, the first step is to demonstrate that the system's properties solve an adaptive problem in a well-engineered way (Dawkins, 1996; Tooby & Cosmides, 1992, 2005; Williams, 1966). But this requires a well-specified theory of the adaptive problem in question. How can one develop such a theory about *social* behavior?

From an evolutionary point of view, the design of programs causing social behavior is constrained by the behavior of other agents. More precisely, it is constrained by the design of the behavior-regulating programs in other agents and the fitness consequences that result from the social interactions these programs cause. These constraints can be analyzed using evolutionary game theory (Maynard Smith, 1982). The application of game theory to the evolutionary process provides a bridge between evolutionary biology and the cognitive sciences that is especially relevant to analyzing social behavior.

An *evolutionarily stable strategy* (ESS) is a strategy (a decision rule) that can persist in a population because it produces fitness outcomes greater than or equal to alternative strategies (Maynard Smith, 1982). The rules of reasoning and decision making that guide social exchange in humans would not exist unless they had outcompeted alternatives, so we should expect that they implement an ESS.[7] By using game theory and conducting computer simulations of the evolutionary process, one can determine which strategies for engaging in social exchange are ESSs.

During the 1960s, evolutionary biologists became very interested in understanding how adaptations causing individuals to help others—often at some cost to themselves—could evolve (Hamilton, 1964; Williams, 1966). To that end, they began exploring interactions that fit the repeated Prisoners' Dilemma, using evolutionary game theory (Axelrod, 1984; Axelrod & Hamilton, 1981; Boyd, 1988; Trivers, 1971). Simulations of the evolutionary process showed that only certain decision rules for conferring benefits on unrelated others were evolutionarily stable. For example, "Always cooperate," a decision rule that distributes benefits to others regardless of whether the recipients ever provide benefits in return, is not evolutionarily stable—it is selected out in any environment that includes decision rules that sometimes "cheat" (fail to reciprocate). But "Tit for tat," a decision rule that helps reciprocators but not cheaters, is an ESS. It can invade a population of cheaters (e.g., individuals equipped with decision rules such as "Always defect," which accepts benefits but never provides benefits in return) and, once established, it can remain at high relative frequencies even when "cheater" designs, such as "Always defect," are introduced into the population.

These analyses showed that ability to reliably and systematically detect cheaters is a necessary condition for cooperation in the repeated Prisoners' Dilemma to be an ESS (and also in other situations; see Stevens & Stephens, 2004; Tooby & Cosmides, 1996). To see this, consider the fate of a program that, because it cannot detect cheaters, bestows benefits on others unconditionally. These unconditional helpers will increase the fitness of any nonreciprocating design they meet in the population. But when a nonreciprocating design is helped, the unconditional helper never recoups the expense of helping: the helper design incurs a net fitness cost while conferring a net fitness advantage on a design that does not help in return.[8] As a result, a population of unconditional helpers is easily invaded and eventually outcompeted by designs that accept the benefits helpers bestow without reciprocating them. Unconditional helping is not an ESS.

In contrast, program designs that cause *conditional* helping—that help those who reciprocate the favor, but not those who fail to reciprocate—can invade a population of nonreciprocators and outcompete them. This is because they gain the benefits of cooperation whenever they interact with another reciprocating design. Moreover, a population of such designs can resist invasion by designs that do not nonreciprocate (cheater designs). Therefore, conditional helping, which requires the ability to detect cheaters, is an ESS. There is a mutual provisioning of benefits, each conditional on the other's compliance.

Engineers always start with a task analysis before considering possible design solutions. We did, too. By applying ESS analyses to the behavioral ecology of hunter-gatherers, we were able to specify tasks that an information-processing program would have to be good at solving for it to implement an evolutionarily stable form of social exchange (Cosmides, 1985, chapter 5; Cosmides & Tooby, 1989). This task analysis of the required computations, *social contract theory*, specifies what counts as good design in this domain. We will mention a few elements of the theory here.

II.i Social Contract Theory

Selection pressures favoring social exchange exist whenever one organism (the provider) can change the behavior of a target organism to the provider's advantage by making the target's receipt of that benefit *conditional* on the target's acting in a required manner. Thus, we define a social exchange as a situation in which, in order to be entitled to receive a *benefit* from another agent, an individual is obligated to satisfy a requirement imposed by that agent (often, but not necessarily, at some cost to himself or herself; but see Stevens & Stephens, 2004; Tooby & Cosmides, 1996). Those who are rationing access to the benefit impose the requirement because its satisfaction creates a situation that benefits them.

A *social contract* expresses this intercontingency and can be expressed in the form of a conditional rule: "If you accept a benefit from agent X, then you must satisfy X's requirement." The agent can be an individual or set of individuals (Cosmides & Tooby, 1989; Fiddick et al., 2000; Tooby, Cosmides, & Price, 2006) and the *must* is understood as involving obligation rather than logical necessity. The specific linguistic expression is not what is important—what matters is that the situation is interpreted as involving one individual offering to provide a benefit to another contingent on that person satisfying a requirement in return, either now or in the future. Intentions to initiate exchange relationships need not be explicitly stated, either: When Agent X provides a benefit to Agent Y, triggering the expectation in both that Y will at some point provide a benefit to X in return, a social exchange relationship has been initiated. Indeed, within hunter-gatherer bands, many or most reciprocity interactions are implicit. For example, in discussing her feelings about food sharing, Nisa, a !Kung San gatherer in Botswana who was extensively interviewed by Marjorie Shostak, explained:

> If a person doesn't give something to me, I won't give anything to that person. If I'm sitting eating, and someone like that comes by, I say, "Uhn, uhn. I'm not going to give any of this to you. When you have food, the things you do with it make me unhappy. If you even once in a while gave me something nice, I would surely give some of this to you." (Shostak, 1981, p. 89)

Nisa's words express her expectations about social exchange, which form an implicit social contract: *If you are to get food in the future from me, then you must share food with me*. Whether we are San foragers or city dwellers, we all realize that the act of accepting a benefit from someone triggers an obligation to behave in a way that somehow benefits the provider, now or in the future.

II.i.i Interpretation by Social Contract Algorithms When the mind registers that a situation with this structure of agents, benefits to agents, and requirements obtains—that is, when the situation is interpreted as involving an implicit or explicit agreement to engage in social exchange—social contract algorithms will apply deontic concepts such as *obligation* and *entitlement*. If the conditional is stated linguistically, they will "read in" these concepts even if they are not stated in the rule. The social contract algorithms will also infer the many implications and entailments of an exchange, listed in tables 2.1 and 2.2. For example, the following two statements will be thought to entail one another:

[1] "If you accept a benefit from agent X, then you are obligated to satisfy X's requirement."

[2] "If you satisfy agent X's requirement, then you are entitled to the benefit X offered to provide."

Because concepts such as *obligation* and *entitlement* will be read in whenever the mind registers that the situation has the deep structure of an exchange, the surface, linguistic structure of the conditional need not contain any of these words. For example, if I know my teen-aged daughter wants to borrow my car, I might say:

[3] "If you borrow my car, then fill the tank with gas" (no modal operators stated).

My daughter, by virtue of having a mind equipped with social contract algorithms, will realize that this implies:

[4] "If I fill her tank with gas, then I will be entitled to borrow her car."

Her mind will supply the concept of *entitlement*, even though I never used the word or even discussed the concept (I only mentioned what I required of *her*). If she fills the tank and I then say she cannot borrow the car after all, she should feel cheated and angry—because I have deprived her of something she now feels entitled to.

Alternatively, I could have said:

[5] "If you fill the tank with gas, then you can borrow my car" (modal *can* stated; automatically interpreted by her social contract algorithms as involving entitlement rather than possibility).

My daughter would understand that this entails:

[6] "If I borrow her car, then I will be obligated to fill her tank with gas" (*obligation* read in by her social contract algorithms).

To import concepts of obligation and entitlement into social exchange in the right way, certain situational cues must be present. Information

Table 2.1

Exchanges: Inferences licensed by social contract algorithms

"If you give me P, then I will give you Q" (= "If I give you Q, then you give me P")* either expression means (entails) the following:

1. I want you to give me *P*,
2. My offer fulfills the cost/benefit requirements of a sincere contract (*listed in table 2.2*),
3. I realize, and I intend that you realize, that 4–9 are entailed if, and only if, you accept my offer:
4. If you give me *P*, then I will give you *Q*,
5. By virtue of my adhering to the conditions of this contract, my belief that you have given (or will give) me *P* will be the cause of my giving you *Q*,
6. If you do not give me *P*, I will not give you *Q*,
7. By virtue of my adhering to the conditions of this contract, my belief that you have not given (or will not give) me *P* will be the cause of my not giving you *Q*,
8. If you accept *Q* from me, then you are obligated to give me *P* (alternatively, If you accept *Q* from me, then I am entitled to receive *P* from you),
9. If you give me *P*, then I am obligated to give you *Q* (alternatively, If you give me *P*, then you are entitled to receive *Q* from me).

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

a. You have agreed to do *P* for me under certain contractual conditions (such as 1–9), and

b. Those conditions have been met, and

c. By virtue of your not thereupon doing *P*, you agree that if I use some means of getting *P* (or its equivalent) from you that does not involve getting your voluntary consent, then I will suffer no reprisal from you. *OR: By virtue of your not thereupon giving me P, you agree that if I lower your utility by some (optimal) amount X (where X >* B_{you}*—your unearned gains), then I will suffer no reprisal from you.*

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

d. I have agreed to give you *Q* under certain contractual conditions (such as 1–9), and

e. Those conditions have not been met, and

f. By virtue of my not thereupon giving you *Q*, I agree that if you use some means of getting *Q* (or its equivalent) from me that does not involve getting my voluntary consent, then you will suffer no reprisal from me. *OR: By virtue of my not thereupon giving you Q, I agree that if you lower my utility by some (optimal) amount X (where X >* B_{me}*—my unearned spoils), then you will suffer no reprisal from me.*

*We used a case of giving as illustration, but social exchange encompasses more than the exchange of goods. "Give" takes three arguments: two agents and the entity given. From this perspective, it is important to preserve the correct binding of agents to items of exchange, and the intercontingent nature of the giving. It is *not* relevant which agent is the subject in the if-clause and which in the then-clause. Furthermore, the entailments all hold, regardless of who fulfills their part of the contract first (i.e., tense is irrelevant, unless it is specified that order falls under the terms of the contract).

Table 2.2
Sincere social contracts: Cost/benefit relations for an exchange of goods when one party is sincere and that party believes the other party is also sincere*

	"If you give me P, then I'll give you Q"			
	Sincere Offer		Sincere Acceptance	
My offer:	*I believe:*		*You believe:*	
P	B_{me}	C_{you}	B_{me}	C_{you}
not-P	0_{me}	0_{you}	0_{me}	0_{you}
Q	C_{me}	B_{you}	C_{me}	B_{you}
not-Q	0_{me}	0_{you}	0_{me}	0_{you}
Profit Margin:	*Positive:* $B_{me} > C_{me}$	*Positive:* $B_{you} > C_{you}$	*Positive:* $B_{me} > C_{me}$	*Positive:* $B_{you} > C_{you}$
Translation:				
My terms	"If B_{me} then C_{me}"		"If B_{me} then C_{me}"	
Your terms	"If C_{you} then B_{you}"		"If C_{you} then B_{you}"	

*Costs and benefits are relative to a baseline that each party believes would pertain in the absence of an exchange (the zero-level utility baseline). B_x = benefit to individual x; C_x = cost to individual x; 0_x = no change from x's utility at baseline. A contract has been *sincerely* offered and accepted when both parties are being truthful about their baselines and when each believes the B > C constraint holds for the other (Cosmides, 1985; Cosmides & Tooby, 1989).

about what each agent wants (and controls access to) is important: here, that my daughter wants to borrow my car and that I would like to have gas in my tank the next time I drive it (i.e., I control access to something she wants, and there is something I want that she can provide). If that is known to the agents (or to an observer), then we need not use "Ifs" and "thens" in speaking. My daughter could say:

[7] "I need to borrow your car"

and I could reply

[8] "OK, but fill the tank with gas."

The structure of entailments in table 2.1 is triggered by the situation of my agreeing ("OK") to provide a benefit contingently ("but do X"). The social contract algorithms supply all the inferences in table 2.1 and inject the appropriate concepts of obligation and entitlement, *whether they were explicitly stated or not.*

We import all this surplus structure so automatically and intuitively that there seems to be nothing to explain: that we interpret the situation in this way is just common sense. True enough—but that is what we are trying to explain. More specifically, we are trying to understand the programming structure of the computational machinery that produces this common sense. Various sets of machinery are involved. Theory-of-mind machinery computes the desires of agents based on cues such as what a person is looking at (Charlie task), moving toward, or saying (Baron-Cohen, 1995). These representations feed into mechanisms designed to detect certain (ancestral) situation types: social exchange, threat, precautions, courtship, and so forth. Social exchange situation detectors are activated when the situation is registered as having the structure of criss-crossing wants and access described above (I control access to what she wants; she can do something I want) plus an indication of agreement to exchange. The cues can be very minimal (see Fiddick et al., 2000, Experiment 1) or floridly expressed as a conditional rule with deontic operators. But if the social exchange situation detectors fire, the social contract algorithms that embody the inferences of table 2.1 will be applied. As a result, you will understand (e.g.) that [3] and [9] entail one another—indeed, express the same social contract—even though neither contains any modal operators, deontic or otherwise:

[3] "If you borrow my car, then fill the tank with gas."

[9] "If you fill the tank with gas, then borrow the car."

Here is the key point, but one must think computationally to understand it. Let's say the computational machinery of our minds contained no social contract algorithms or other domain-specific inference systems. Let's say the only inferential rules implemented by our computational machinery were those of first-order logic, and that these rules were designed, as is common in programming languages, to recognize and operate over antecedents (*P*s) and consequents (*Q*s) as explicitly stated in a conditional rule, *If P then Q*. If this were the case, our common sense would *not* tell us that [3] entails [9] and vice versa. By the rules of inference of first-order logic, sentences [3] and [9] are not logically equivalent: *If P then Q* does not imply *If Q then P*.

Now imagine that the human mind contains computational machinery that implements social contract theory in the way we have proposed. On this view, we would see [3] as implying [9] because (i) social contract algorithms supply concepts of *obligation* and *entitlement* in just the right places when situation detectors have registered that the criss-crossing pattern of access and wants applies and an agreement to provide benefits contingently has been reached, and (ii) they apply a set of inference rules *specific* *to the domain of social exchange* that license each as a translation of the other (No. 8 and No. 9 in table 2.1). This will happen because, when the situation detector has activated the social contract algorithms, they will map [3] onto the deep structure specified in [1], and [9] onto the deep structure specified in [2].

II.i.i.i Applying Multiple Systems What happens if there are multiple sets of computational machinery in the mind, each implementing different rules of inference? Each would have to be equipped with situation detectors, which monitor for cues indicating whether the current situation fits the input conditions for a given adaptive specialization (Barrett, 2005; Sperber, 1994). Many situations will activate only one domain-specific inference system, but some might jointly activate two of them. For example, Cheng and Holyoak (1985) used an international border-crossing rule indicating that "If you are to enter the country, then you must have been vaccinated for cholera." This is clearly a social contract for the person who very much *wants* to enter the country. But for someone who views entering the country as hazardous because cholera is endemic—and for government officials who want to stem the tide of cholera in the country—this may be interpreted as a precautionary rule, fitting the template "If you are to engage in hazardous activity H, then take precaution R." We have also suggested that a well-designed mind containing many different inference systems may implement a *principle of preemptive specificity* (Fiddick et al., 2000). According to this metaprinciple, if a situation fits the input conditions of several inference systems in a class inclusion relationship, the most domain-specific one will preempt the operation of the more domain-general ones. For example, if, in addition to social contract algorithms, the mind also contains computational machinery that implements *modus ponens* and *modus tollens*, these will not be applied to [3] and [9]. When the situation is interpreted as involving exchange, the social contract algorithms will preempt the operation of these more domain-general inference rules. This predicts, for example, that it will be difficult to apply *modus ponens* and *tollens* to [9]. This turns out to be true, as we will explain in section IV.iii.i.

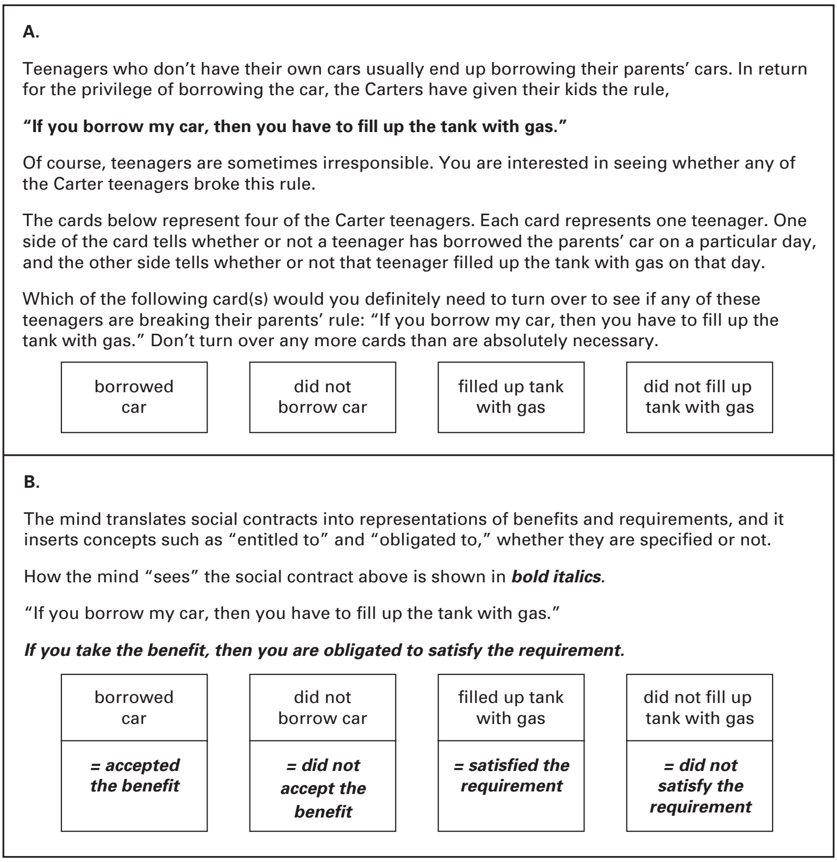
II.i.i.ii Surplus Structure Is Domain Specific We have discussed the surplus structure of obligation, entitlement, and inference that social contract algorithms supply to rules like [3] and [9]. But note that none of this happens for indicative rules. We do not interpret the Ebbinghaus rule as meaning "If a person has Ebbinghaus disease, then that person is *obligated* to be forgetful," nor do we think it implies that "If a person is forgetful, then that person is *entitled* to have Ebbinghaus disease."

Like social contracts, precautionary rules are considered deontic: they specify what you *ought* to do (when engaged in hazardous activities). Precautionary rules also activate surplus structure, but it is different from that applied to social contract rules. The *must* in "If you work with toxic gases, then you must wear a gas mask" does not imply that one has *incurred an obligation* to another person by virtue of working with toxic gases. Instead it expresses advice about how to stay safe while working with toxic gases (i.e., "You must wear it or you will be in danger").

Moreover, it would be bizarre to infer from this rule that "If you wear a gas mask, then you are *entitled* to work with toxic gas." Few people view working with toxic gases as a benefit. Having put on the mask, most people would be relieved, not angry, if I said that they don't have to work with toxic gases after all. Switching the order of the clauses within the If-then structure of the rule should cause the precautionary inference system to assign the following interpretation: "If you wear a gas mask, then *it will be safe* for you to work with toxic gases" (Fiddick et al., 2000).

What about someone who really *wants* to work with toxic gases, that is, someone who views doing so as a *benefit* that I control access to and am then denying them, despite their having met my requirement? A person with these desires would be interpreting the toxic gas rule as a social contract: "If you want the benefit (of working with toxic gases), then you must satisfy my requirement." Entitlement would then be activated by the social contract algorithms, as well as anger when that entitlement is denied. That is, *not all deontic rules activate the same surplus structure*. If situation detectors map it onto the template of a social contract ([1] or [2]), *entitlement* and *obligation* will be imported in the appropriate places. But these concepts will not be imported if situation detectors map the rule onto the hazard-precaution template of a precautionary rule. In that case, words such as "must" or "ought" will be interpreted as referring to the causal requirements of safety (section IV.iv.).

II.i.ii Cheater Detection and Social Exchange Let us now turn from interpretation to cheater detection. Figure 2.2a shows a Wason selection task in which the conditional is the social contract rule about borrowing the car discussed above. In contrast to the Ebbinghaus disease rule (figure 2.1), which elicited "*P* & *not-Q*" responses from only 26% of undergraduates, this social contract rule elicited that response from 76% of undergraduates. This is logically correct performance. But our claim is not that social



A.

Teenagers who don't have their own cars usually end up borrowing their parents' cars. In return for the privilege of borrowing the car, the Carters have given their kids the rule,

"If you borrow my car, then you have to fill up the tank with gas."

Of course, teenagers are sometimes irresponsible. You are interested in seeing whether any of the Carter teenagers broke this rule.

The cards below represent four of the Carter teenagers. Each card represents one teenager. One side of the card tells whether or not a teenager has borrowed the parents' car on a particular day, and the other side tells whether or not that teenager filled up the tank with gas on that day.

Which of the following card(s) would you definitely need to turn over to see if any of these teenagers are breaking their parents' rule: "If you borrow my car, then you have to fill up the tank with gas." Don't turn over any more cards than are absolutely necessary.

borrowed car	did not borrow car	filled up tank with gas	did not fill up tank with gas

B.

The mind translates social contracts into representations of benefits and requirements, and it inserts concepts such as "entitled to" and "obligated to," whether they are specified or not.

How the mind "sees" the social contract above is shown in ***bold italics.***

"If you borrow my car, then you have to fill up the tank with gas."

If you take the benefit, then you are obligated to satisfy the requirement.

borrowed car	did not borrow car	filled up tank with gas	did not fill up tank with gas
= accepted the benefit	***= did not accept the benefit***	***= satisfied the requirement***	***= did not satisfy the requirement***

Figure 2.2
Wason task with a social contract rule. (A) In response to this social contract problem, 76% of subjects chose *P* & *not*-*Q* ("borrowed the car" and "did not fill the tank with gas")—the cards that represent potential cheaters. Yet only 26% chose this (logically correct) answer in response to the descriptive rule in figure 2.1. Although this social contract rule involves familiar items, unfamiliar social contracts elicit the same high performance. (B) How the mind represents the social contract shown in (A). According to inferential rules specialized for social exchange (but not according to first-order logic), "If you take the benefit, then you are obligated to satisfy the requirement" implies "If you satisfy the requirement, then you are entitled to take the benefit." Consequently, the rule in (A) implies: "If you fill the tank with gas, then you may borrow the car" (see figure 2.4, switched social contracts).

contract content activates computational machinery implementing first-order logic. Our claim is that it activates social contract algorithms, which interpret the rule as explained above, and which contain a subroutine for detecting cheaters: individuals who cheat by design, not by accident. The function of the cheater detection mechanism is to direct information search in a way that will uncover potential cheaters. Searching for violations requires more than having semantic knowledge of what counts as a violation;[9] as we discussed above, people know that the combination of *P* and *not-Q* violates indicative rules, but they do not spontaneously *search* for information that could reveal potential violations, *even when they are explicitly asked to do so*. By extension, searching for cheaters requires more, computationally, than having a mental representation of what counts as cheating. We have therefore posited an information search routine that directs attention to anyone who has taken the benefit (have they satisfied the requirement?) and to anyone who has not satisfied the requirement (have they taken the benefit?). By coincidence, those cards map onto the logically correct response, "*P & not-Q*" for the rule as stated in figure 2.2a. Figure 2.2b shows our view of how the mind sees this problem. According to social contract theory, social contract algorithms are not mapping propositions in the conditional onto a syntax of true antecedents and false consequents; they are mapping them onto representations of an agent who has *accepted the benefit* and of an agent who has *not satisfied the requirement*, and the cheater detection mechanism is directing information search to those agents. If the cards representing those two conditions happen to correspond to *P* and *not-Q*, it will look like people reasoned in accordance with first-order logic. However, it is easy to create problems in which looking for cheaters will produce a response that violates first-order logic, as we will show below.

According to social contract theory, cheating does involve the violation of a conditional rule, but it is a particular *kind* of violation of a particular *kind* of conditional rule. The rule must fit the template for a *social contract* ([1] or [2]); the violation must be one in which an individual *intentionally* took what *that* individual considered to be a *benefit* and did so without satisfying the requirement.

III The Design of Computational Machinery That Governs Reasoning about Social Exchange

Because social contract theory provides a standard of good design against which human performance can be measured, there can be a meaningful answer to the question, “Are the programs that cause reasoning about social exchange well engineered for the task?” Well-designed programs for engaging in social exchange—if such exist—should include features that execute the computational requirements specified by social contract theory and do so reliably, precisely, and economically (Williams, 1966).

From social contract theory’s task analyses, we derived a set of predictions about the design features that a neurocognitive system specialized for reasoning about social exchange should have (Cosmides, 1985; Cosmides & Tooby, 1989). The following six design features (D1–D6) were among those on the list:

D1. Social exchange is cooperation for mutual *benefit*. If there is nothing in a conditional rule that can be interpreted as a rationed benefit, then interpretive procedures should not categorize that rule as a social contract. To trigger the inferences about obligations and entitlements that are appropriate to social contracts, the rule must be interpreted as restricting access to a benefit to those who have met a requirement. (This is a necessary, but not sufficient, condition; Cosmides & Tooby, 1989; Gigerenzer & Hug, 1992.)

D2. Cheating is a specific way of violating a social contract: It is taking the benefit when you are not entitled to do so. Consequently, the cognitive architecture must define the concept of *cheating* using contentful representational primitives, referring to illicitly taken *benefits*. This implies that a system designed for cheater detection will not know what to look for if the rule specifies no benefit to the potential violator.

D3. The definition of cheating also depends on which agent’s point of view is taken. Perspective matters because the item, action, or state of affairs that one party views as a benefit is viewed as a requirement by the other party. The system needs to be able to compute a cost/benefit representation from the perspective of each participant and define cheating with respect to that perspective-relative representation.

D4. To be an ESS, a design for conditional helping must not be outcompeted by alternative *designs*. Accidents and innocent mistakes that result in an individual’s being cheated are not markers of a design difference. Moreover, decision rules that punish accidental violations of social contracts by refusing to cooperate further with the accidental violator lose many opportunities to gain from cooperation; simulation results show such strategies get selected out in the presence of strategies that exclude only intentional cheaters (Panchanathan & Boyd, 2003). A cheater detection system should look for cheat*ers*: individuals equipped with programs that cheat by design.[10] Hence, intentional cheating should powerfully trigger the detection system, whereas mistakes should trigger it weakly or not at all. (Mistakes that result in an individual's being cheated are relevant only insofar as they may not be true mistakes.)

D5. The hypothesis that the ability to reason about social exchange is acquired through the operation of some general-purpose learning ability necessarily predicts that good performance should be a function of experience and familiarity. In contrast, an evolved system for social exchange should be designed to recognize and reason about social exchange interactions no matter how unfamiliar the interaction may be, provided it can be mapped onto the abstract structure of a social contract. Individuals need to be able to reason about each new exchange situation as it arises, so rules that fit the template of a social contract should elicit high levels of cheater detection, even if they are unfamiliar.

D6. Inferences made about social contracts should not follow the content-free rules of first-order logic. They should follow a content-specific adaptive logic, evolutionarily tailored for the domain of social exchange, as presented in the discussion of interpretation above (Cosmides, 1985; Cosmides & Tooby, 1989).

Each of these design features has now been tested for and empirically validated. The mechanisms that govern reasoning about social exchange have many improbable properties that are well engineered for solving adaptive problems arising in this domain. Various hypotheses positing machinery designed to operate over a class of content more general than social exchange have been proposed, but none are capable of explaining the pattern of results found. Not only is the computational machinery specialized but the process for its developmental acquisition appears to be specialized as well: the developmental evidence of precocial competence and the distribution of reasoning specializations that characterize the adult state are not consistent with any domain-general proposals for their developmental acquisition (for review and discussion, see Cosmides & Tooby, 2005a).

Cosmides and Tooby (2005a) review the design evidence that supports the claim that the human mind reliably develops an adaptive specialization for reasoning about social exchange and that rules out by-product hypotheses. In this chapter, we focus on how the evidence bears on two somewhat different issues: (i) Can this evidence be explained by positing a "deontic logic machine," that is, computational machinery that implements a domain-general form of deontic logic? (ii) If not, then what are the prospects for creating a content-general deontic logic that still respects the empirical facts about deontic reasoning?

IV Conditional Reasoning and Social Exchange: Some Empirical Findings

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

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

As discussed earlier, the Wason selection task is a standard tool for investigating conditional reasoning. Because it asks subjects to look for potential violations of conditional rules, it was particularly well suited to our purposes: We were interested in cheater detection, which is a (specialized) form of violation detection. Using this task, an extensive series of experiments has now been conducted that addresses the following questions:

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

If the human brain had cognitive machinery that causes good conditional reasoning regardless of content, then people should be good at tasks requiring conditional reasoning. For example, they should be good at detecting violations of indicative conditional rules. Yet studies with the Wason selection task showed that they are not. If our minds were equipped with reasoning procedures specialized for detecting *logical* violations of conditional rules, the correct answer (choose *P*, choose *not-Q*) would be intuitively obvious and pop out for people. But it is not.

First-order logic provides a standard of good design for content-general conditional reasoning: its inference rules were constructed by philosophers to generate true conclusions from true premises, regardless of the subject matter one is asked to reason about. When human performance is measured against this standard, there is little evidence of good design: Conditional rules with descriptive (indicative) content fail to elicit logically correct performance from 70% to 95% of people, even when the content involves familiar terms drawn from everyday life (Cosmides, 1989; Griggs & Cox, 1982; Wason, 1983; Manktelow & Evans, 1979; Sugiyama et al., 2002). Therefore, one can reject the hypothesis that the human mind is equipped with reasoning machinery that causes good violation detection on all conditional rules, regardless of their content or domain.

IV.i A Dissociation by Content (D1, D2)

People are poor at detecting violations of conditional rules when their content is descriptive. But this result does not generalize to conditional rules that express a social contract. People who ordinarily cannot detect violations of if-then rules can do so easily and accurately when that violation represents cheating in a situation of social exchange. This pattern—good violation detection for social contracts but not for descriptive rules—is a dissociation in reasoning elicited by differences in the conditional rule's *content*. It provides (initial) evidence that the mind has reasoning procedures specialized for detecting cheaters.

The high performance—76% correct—found for the borrowing car rule of figure 2.2 is not particular to that problem: it is found whenever subjects are asked to look for violations of a conditional rule that fits the social contract template—"If you take benefit B, then you must satisfy requirement R." For *standard* social contracts (ones with the benefit to the cheater in the antecedent clause), people check the individual who accepted the benefit (*P*) and the individual who did not satisfy the requirement (*not-Q*). These are the cases that represent potential cheaters (figure 2.2b). The adaptively correct answer is immediately obvious to most subjects, who commonly experience a pop-out effect. No formal training is needed. Whenever the content of a problem asks one to look for cheaters in a social exchange, subjects experience the problem as simple to solve, and their performance jumps dramatically. In general, 65% to 80% of subjects get it right, the highest performance found for a task of this kind (for reviews, see Cosmides, 1985, 1989; Cosmides & Tooby, 1992, 1997; Fiddick et al., 2000; Gigerenzer & Hug, 1992; Platt & Griggs, 1993).

The content-blind syntax of first-order logic would treat investigating the person who borrowed the car (*P*) and the person who did not fill the gas tank (*not-Q*) as logically equivalent to investigating the person with Ebbinghaus disease (*P*) and the person who is not forgetful (*not-Q*) for the Ebbinghaus problem in figure 2.1. But everywhere it has been tested (adults in the United States, United Kingdom, Germany, Italy, France, Hong Kong, Japan; schoolchildren in Quito, Ecuador; Shiwiar hunter-horticulturalists in the Ecuadorian Amazon), people do not treat social exchange problems as equivalent to other kinds of conditional reasoning problems (Cheng & Holyoak, 1985; Cosmides, 1989; Hasegawa & Hiraishi, 2000; Platt & Griggs, 1993; Sugiyama, Tooby, & Cosmides, 2002; supports D5, D6). Their minds distinguish social exchange content from other domains and reason as if they were translating their terms into representational primitives such as *benefit*, *cost*, *obligation*, *entitlement*, *intentional*, and *agent* (figure 2.2b; Cosmides & Tooby, 1992, 2005a; Fiddick et al., 2000). Reasoning problems could be sorted into indefinitely many categories based on their content or structure (including first-order logic's two content-free categories, antecedent and consequent). Yet, even in remarkably different cultures, the same mental categorization occurs. This cross-culturally recurrent dissociation by content was predicted in advance of its discovery by social contract theory's adaptationist analysis.

It is worth noting that cognitive disorders can impair logical reasoning abilities without affecting the ability to detect cheaters on social contracts. Maljkovic (1987) compared a group of patients with schizophrenia to a control group with no cognitive impairment. She found that the schizophrenic patients had deficits compared to the controls on a battery of (non-Wason) logical reasoning tasks, in a way consistent with frontal lobe dysfunction. However, the schizophrenic patients did well—indeed just as well as the control subjects—on Wason tasks where the rule was a social contract and a violation was cheating. That their social contract reasoning was intact is striking because individuals with schizophrenia typically manifest deficits on virtually any test of general intellectual functioning they are given (McKenna, Clare, & Baddeley, 1995). Maljkovic's interesting result is consistent with the view that social contract reasoning is accomplished by a dedicated system that is functionally dissociable from more general forms of logical reasoning.

IV.ii Do Unfamiliar Social Contracts Elicit Cheater Detection? (D5)

An individual needs to understand each new opportunity to exchange as it arises, so it was predicted that social exchange reasoning should operate even for unfamiliar social contract rules (D5). This distinguishes social contract theory strongly from theories that explain reasoning performance as the product of general-learning strategies plus experience: the most natural prediction for such skill-acquisition theories is that performance should be a function of familiarity.

The evidence supports social contract theory: Cheater detection occurs even when the social contract is wildly unfamiliar (figure 2.3a). For example, the rule "If a man eats cassava root, then he must have a tattoo on his face" can be made to fit the social contract template by explaining that the people involved consider eating cassava root to be a benefit (the rule then implies that having a tattoo is the requirement an individual must satisfy to be eligible for that benefit). When given this context, this outlandish, culturally alien rule elicits the same high level of cheater detection as highly familiar social exchange rules. This surprising result has been replicated for many different unfamiliar rules (Cosmides, 1985, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992; Platt & Griggs, 1993). It supports the psychological reality of the template (see [1]) posited for the representation of social exchanges (D1, D2).

IV.ii.i Not Familiarity, Not Memory Retrieval This means the dissociation by content—good performance for social contract rules but not for descriptive ones—has nothing to do with the familiarity of the rules tested. Nor does it reflect the subject's ability to retrieve real-life instances in which they had violated the rule (a hypothesis that was once considered to explain results for the drinking age law, which is a social contract). Familiarity is neither necessary nor sufficient for eliciting high performance.

First, as discussed above, familiarity does not produce high levels of performance for descriptive (indicative) rules. If familiarity fails to elicit high performance on descriptive rules, then it also fails as an explanation for high performance on social contracts. Second, the fact that unfamiliar social contracts elicit high performance shows that familiarity is not necessary for eliciting violation detection. Third (and most surprising), people are just as good at detecting cheaters on culturally unfamiliar or imaginary social contracts as they are for ones that are completely familiar: Cosmides (1985, pp. 244–250) found that performance on unfamiliar social contracts (such as the cassava root one) was just as high as performance on familiar social contracts, including the drinking age problem. This provides a challenge for any counterhypothesis resting on a general-learning skill acquisition account (most of which rely on familiarity and repetition), or on the ability to retrieve violating instances from memory.

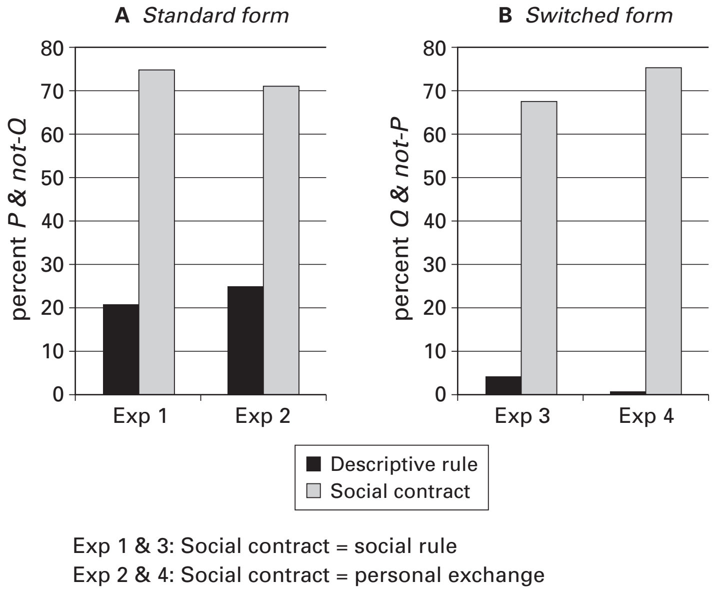


Figure 2.3
Detecting violations of unfamiliar conditional rules: Social contracts versus descriptive indicative rules. In these experiments, the same, unfamiliar rule was embedded either in a story that caused it to be interpreted as a social contract or in a story that caused it to be interpreted as a rule describing some state of the world. For social contracts, the correct answer is always to pick the *benefit accepted* card and the *requirement not satisfied* card. (A) For standard social contracts, these correspond to the logical categories *P* and *not-Q*. *P* & *not-Q* also happens to be the logically correct answer. Over 70% of subjects chose these cards for the social contracts, but fewer than 25% chose them for the matching descriptive rules. (B) For switched social contracts, the *benefit accepted* and *requirement not satisfied* cards correspond to the logical categories *Q* and *not-P*. This is not a logically correct response. Nevertheless, about 70% of subjects chose it for the social contracts; virtually no one chose it for the matching descriptive rule (see figure 2.4).

IV.iii Adaptive Logic, Not First-Order Logic (D3, D6)

The above shows that it is possible to construct social exchange problems that will elicit a logically correct answer. But this is not because social exchange content activates computational machinery implementing first-order logic.

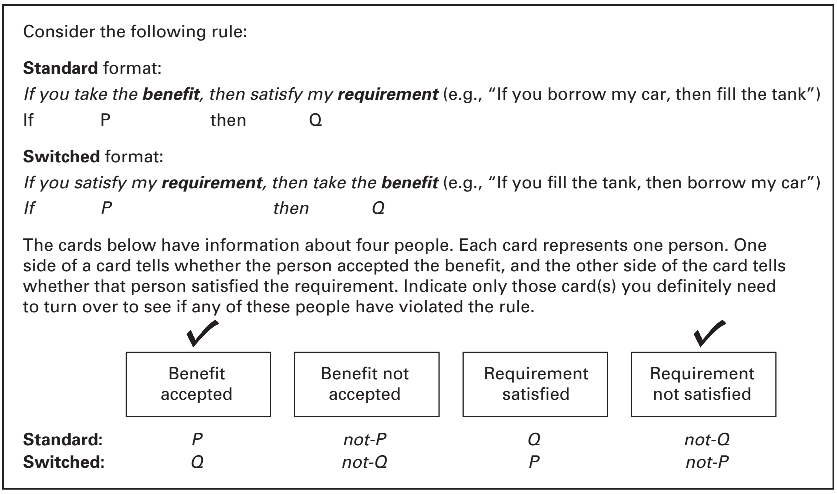
First-order logic is content blind. Computational machinery implementing it—let's call this a "first-order logic machine"—would apply the same definition of violation (co-occurrence of a true antecedent with a false consequent) to all conditional rules, whether they are social contracts, threats, or descriptions of how the world works. This is because (i) first-order logic operates over propositions, regardless of their content, and (ii) it contains no procedures for importing surplus concepts such as *obligation*, *entitlement*, *agent*, *perspective*, or *intention to violate* into the rule. But the definition of cheating implied by design features D1 through D4 does not map onto this content-blind definition of violation. What counts as cheating in social exchange is so content sensitive that it is easy to create problems in which the search for cheaters will result in a logically incorrect response (and the search for logical violations will fail to detect cheaters; see figure 2.4). When given such problems, people look for cheaters, thereby giving a logically incorrect answer (*Q* and *not-P*). Two sets of results bear on this issue: experiments with switched social contracts and perspective change experiments.

IV.iii.i Switched Social Contracts A first-order logic machine would never infer that *If P then Q* implies *If Q then P*. In contrast, social contract algorithms should have domain-specialized deontic rules of inference according to which

[1] "If you accept a benefit from agent X, then you are obligated to satisfy X's requirement," implies

[2] "If you satisfy agent X's requirement, then you are entitled to the benefit X offered to provide"—and vice versa.

A "standard" social contract has the benefit to the potential cheater in the antecedent (*P*) clause (as in [1]); a "switched" social contract has that benefit in the consequent (*Q*) clause (as in [2]). These terms do not refer to different kinds of social contracts, only to different linguistic expressions of the same offer to exchange. For example, "If you give me your watch, I'll give you $100" and "If I give you $100, then give me your watch" are recognized by normal human minds as expressing the same offer, entailing the same entitlements and obligations for both agents.



Consider the following rule:

Standard format:

*If you take the **benefit**, then satisfy my **requirement*** (e.g., "If you borrow my car, then fill the tank")
If P then Q

Switched format:

*If you satisfy my **requirement**, then take the **benefit*** (e.g., "If you fill the tank, then borrow my car")
If P then Q

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

	✓			✓
	Benefit accepted	Benefit not accepted	Requirement satisfied	Requirement not satisfied
Standard:	*P*	*not-P*	*Q*	*not-Q*
Switched:	*Q*	*not-Q*	*P*	*not-P*

Figure 2.4

Generic structure of a Wason task when the conditional rule is a social contract. A social contract can be translated into either social contract terms (benefits and requirements) or logical terms (*P*s and *Q*s). Check marks indicate the correct card choices if one is looking for cheaters—these should be chosen by a cheater detection subroutine, whether the exchange was expressed in a standard or switched format. This results in a logically incorrect answer (*Q* & *not-P*) when the rule is expressed in the switched format, and a logically correct answer (*P* & *not-Q*) when the rule is expressed in the standard format. By testing switched social contracts, one can see that the reasoning procedures activated cause one to detect cheaters, not logical violations (see figure 2.3B). Note that a logically correct response to a switched social contract—where *P* = *requirement satisfied* and *not-Q* = *benefit not accepted*—would fail to detect cheaters.

Likewise, given the context of criss-crossing desires and access expressed in the car task of figure 2.2, it shouldn't matter whether the rule is expressed in standard form (as in [3]) or as a switched social contract—even one lacking explicit deontic operators, like [9] "If you fill the tank with gas, then borrow the car" (see section II.i.i). Because deontic inferences specialized for exchange will recognize them as implying the same set of obligations and entitlements, the same event will count as cheating by a teenager: borrowing the car without filling the tank.

This leads to a prediction: it should not matter whether the rule is expressed in standard ([3], [1]) or switched ([9], [2]) form. A subroutine for detecting cheaters should check teenagers who have taken the benefit (borrowed the car) and teenagers who have not satisfied the requirement that provision of this benefit was made contingent upon (didn't fill the tank).

By testing switched social contracts, we can discriminate the predictions of first-order logic from those of social contract theory. As figure 2.4 shows, the cheater-relevant cards correspond to the logical categories *P* and *not-Q* for standard social contracts. The cheater detection subroutine will therefore produce the same answer as a first-order logic machine in response to standard social contracts—not because this response is logically correct but because it will correctly detect cheaters. But for switched social contracts, a cheater detection subroutine will not produce a logically correct answer.

In the switched format, the cheater relevant cards, "borrowed the car" and "did not fill tank," correspond to the logical categories *Q* (true consequent) and *not-P* (false antecedent), respectively. Choosing them will correctly detect teenagers who are cheating, but it violates first-order logic: instances of *not-P* cannot violate the material conditional, whether paired with *Q* or *not-Q*. (Please note that social contracts do not imply a biconditional interpretation: filling the tank without borrowing the car may be altruistic, but it is not cheating or any other kind of violation.)

When given switched social contracts with structures like [9] or [2], subjects overwhelmingly respond by choosing *Q* & *not-P*, a logically incorrect answer that correctly detects cheaters (figure 2.3b; Cosmides, 1985, 1989; Gigerenzer & Hug, 1992; Platt & Griggs, 1993; supports D2, D6). They do this even when there are no deontic operators specified in the rule. Indeed, when subjects' choices are classified by logical category, it looks like standard and switched social contracts elicit different responses. But when their choices are classified by *social contract* category, they are invariant: for both rule formats, people choose the cards that represent an agent who took the benefit and an agent who did not meet the requirement.

The hypothesis that social contract content merely activates first-order logic is eliminated by this finding. A first-order logic machine is equipped with only one definition of violation for a conditional rule: *P* & *not-Q*. It would therefore choose those cards in response to switched social contracts like rule [9], even though this would not detect cheaters.[11] A teenager who has filled the tank (*P*) without borrowing the car (*not-Q*) may be altruistic but has not cheated.[12]

Interestingly, Gigerenzer and Hug (1992) had a handful of subjects who answered "*P* & *not-Q*" to every conditional they got. In debriefing, these individuals said they had applied formal logic to all problems—but that this had been particularly difficult for some of them. These turned out to be the switched social contracts. This kind of difficulty is expected: to apply first-order logic, they would have to suppress the spontaneous and intuitive inferences generated by their social contract algorithms (see discussion of the principle of preemptive specificity in section II.i.i.i).

IV.iii.i.i Reasoning in the Rainforest This pattern—choosing the *benefit accepted* and *requirement not met* cards on standard and switched social contracts—is not particular to people raised in advanced market economies. Sugiyama, Tooby and Cosmides (2002) adapted Wason tasks for nonliterate individuals and administered them to Shiwiar hunter-horticulturalists living in a remote part of the Ecuadorian Amazon. Shiwiar subjects were given a standard social contract (like the cassava-tattoo problem), a switched social contract ("If you bring me a basket of fish, then you can borrow my motorboat"), and a descriptive problem. The Shiwiar were just as good at detecting cheaters on Wason tasks as Harvard undergraduates were. For cheater-relevant cards, the performance of Shiwiar hunter-horticulturalists was identical to that of Harvard students: *benefit accepted* and *requirement not met* were chosen by over 80% of subjects.

The Shiwiar, whose way of life involves frequent sharing, differed from Harvard students only in that they were more likely to also show interest in cheater-irrelevant cards—the ones that could reveal acts of generosity. The Shiwiar's excellence at cheater detection did not result from indiscriminate interest in all cards, however. This could be seen by comparing performance on standard and switched social contracts while controlling for logical category (e.g., *P* is cheater-relevant for a standard social contract (where it represents a *benefit accepted*) but cheater-irrelevant for a switched one (where it represents a *requirement satisfied*)). Controlling for logical category, Shiwiar were more than twice as likely to choose a card when it was cheater-relevant than when it was not ($p < .005$).

That there was no difference between cultures in choosing the cheater-relevant cards is expected: For social exchange to implement an ESS, the development of the cheater detection subroutine would need to be buffered against (evolutionarily normal) variations in local cultures. The only "cultural dissociation" we found was in ESS-irrelevant aspects of performance (interest in generosity).

IV.iii.i.ii Deontic Logic? Is choosing *Q* and *not-P* on a switched social contract consistent with content-general deontic logics? Manktelow and Over (1987) considered this question and concluded that the answer is indeterminate. Consider "If you fill the tank with gas, then you may borrow the car." Let's assume the mind contains computational machinery implementing a *domain-general* deontic logic: a deontic logic machine. Let's assume this deontic logic machine interprets this rule as having the deep structure of [2], and therefore meaning "If you fill the tank with gas, then you are *entitled* to borrow the car." Technically, no events could violate this rule: by most deontic logics, the fact that you are *entitled* to borrow the car does not mean that you are *required* to borrow the car. Thus, when subjects are asked to look for violations of the rule, the deontic logic machine should choose no cards at all. But this rarely happens.

For "*Q* & *not-P*" to be chosen, one needs additional assumptions. The instruction to see whether any of the *teenagers* have violated the rule would have to be a cue used by the deontic logic machine. In response to that cue, it would have to derive an *obligation* that falls on *teenagers* by virtue of the rule. To do this, the cue would have to trigger the inference that this switched social contract (with the format of [2]) implies another rule with the format of [1]. Only by that transformation could the subject tell whether the rule implies *any* obligation on the part of car-borrowing teenagers. So, even if the parents' statement had been floridly deontic, such as "If you fill the tank with gas, then you are entitled to borrow the car," the deontic logic machine would have to first derive what this implies about obligations incurred by teenagers. Without doing this, the deontic logic machine could not determine whether a violation has occurred, that is, whether an obligation has remained unfulfilled—which is necessary for choosing the "borrowed car" card (*Q*) and the "didn't fill tank" card (*not-P*).

Social contract algorithms are designed to make translations of precisely this kind: they are hypothesized to be sensitive to which party is the potential rule violator, to have procedures for inferring that the entitlements of that agent specified in [2] imply that that agent has the obligations specified by [1], and to search for information that could reveal whether that agent has cheated. The questions for deontic logicians are: (i) Can one build a deontic logic that does all of these things *without* it being a notational variant of social contract theory? And (ii) is it possible to build a deontic logic that does all of this *and* is general to all deontic rules, whether they are social contracts or not? We suspect the answer to both is no.

In particular, it is not clear how a deontic logic could be general to all (deontic) domains yet license the inference that

[10] "If you satisfy requirement R, then you are entitled to E" implies

[11] "If you get E, then you are obligated to satisfy requirement R" (and vice versa).

These two rules are interesting because they are only *slightly* more domain-general statements of [2] and [1] [10] subsumes the more domain-specific [2] as a special case; [11] similarly subsumes [1]). However, [10] and [11] cannot imply one another *across domains* without violating our moral intuitions. For example,

"If you are an American citizen, then you are entitled to be free from torture" (instance of [10])

is not usually thought to imply

"If you are free from torture, then you are *obligated* to be an American citizen" (instance of [11]).

In contrast, when the entitlement E is a benefit contingently offered *in the context of social exchange*, [2] entails [1] and vice versa. That is, the conditions under which an inference from entitlement to obligation is valid are quite domain-specific—they do not encompass all moral domains.

IV.iii.ii Perspective Change As predicted (D3), the mind's automatically deployed definition of cheating is tied to the perspective you are taking (Gigerenzer & Hug, 1992). For example, consider the following social contract:

[12] "If an employee is to get a pension, then that employee must have worked for the firm for over ten years."

This social contract rule elicits different answers depending on whether subjects are cued into the role of employer or employee. Those in the employer role look for cheating by employees, investigating cases of *P* and *not-Q* (employees with pensions; employees who have worked for fewer than ten years). Those in the employee role look for cheating by employers, investigating cases of *not-P* and Q (employees with no pension; employees who have worked more than ten years). *Not-P* & *Q* is correct if the goal is to find out whether the employer is cheating employees. But it is not *logically* correct by first-order logic, which, because it is content-blind, has no role for agents and their differing perspectives.[13]

In this social exchange, the benefit to one agent is the requirement for the other: giving pensions to employees benefits the employees but is the requirement the employer must satisfy (in exchange for the benefit to the employer of more than ten years of employee service). To capture this distinction between the perspectives of the two agents, rules of inference for social exchange must be content sensitive, defining benefits and requirements relative to the agents involved. Because the rules of first-order logic are blind to the content of the propositions over which they operate, they have no way of representing the values of an action to each agent involved.

Can a domain-general deontic logic do the trick? The answer is the same as for the switched social contracts. To determine whether the employ*er* has cheated, the deontic logic machine would have to determine what obligation, if any, the employer owes to the employee. But rule [12] specifies no such obligation. To derive one, the deontic logic machine would need procedures that take [12] as input and derive [13] from it:

[13] "If the employee has worked for the firm for more than ten years, then the employer is obligated to give that employee a pension."

But what justifies this inference? Social contract algorithms derive this inference when they recognize that a long-working employee is a *benefit* to the employer, for which the employer is willing to give pensions. That is, [13] is derived from a representation something like this:

[14] "If the employer gets the *benefit from the employee* (of long service), then the employer is obligated to give a benefit (the pension) to that employee."

It is difficult to see how a content-blind deontic logic machine could perform this transformation as it depends on underlying representations of *benefits* to agents (and remember, the more general formulation in section IV.iii.i.ii won't work). Again, we need to ask whether there can be a version of deontic logic that is not merely a notational variant of social contract theory. Moreover, if one constructs a deontic logic that does implement the same rules as social contract algorithms, then we have to ask how it is going to achieve its content generality. How will it operate over precautionary rules, rules of etiquette, and other deontic conditionals that lack the structure of social exchange?

IV.iv How Many Specializations for Conditional Reasoning?

Social contracts are not the only conditional rules for which natural selection should have designed specialized reasoning mechanisms (Cosmides, 1989). Indeed, good violation detection is also found for conditional rules drawn from two other domains: threats (Tooby & Cosmides, 1989) and precautionary rules. Threats are not deontic, but precautionary rules are. So we need to ask whether good performance on social contracts and precautionary rules is caused by a single neurocognitive system or by two functionally distinct ones. We also need to ask whether all deontic rules elicit good violation detection. If the answer is no, then reasoning about social exchange cannot be caused by a deontic logic machine that operates in a uniform manner on all deontic conditionals.

IV.iv.i Precautionary Rules Game theory is not needed to see that there would be a fitness advantage to machinery that makes one good at detecting when someone is in danger by virtue of having violated a precautionary rule. Precautionary rules are ones represented as fitting the template: "*If one is to engage in hazardous activity H, then one must take precaution R*" (e.g., "If you are working with toxic gases, then wear a gas mask"). Using the Wason task, it has been shown that people are very good at detecting potential violators of precautionary rules, that is, individuals who have engaged in a hazardous activity without taking the appropriate precaution (e.g., those working with toxic gases (*P*) and those not wearing a gas mask (*not-Q*)). Indeed, relative to descriptive rules, precautions show a spike in performance, and the magnitude of this content effect is about the same as that for detecting cheaters on social contracts (Cheng & Holyoak, 1989; Fiddick et al., 2000; Manktelow & Over, 1988, 1990; Stone et al., 2002; Ermer, Guerin, Cosmides, Tooby, & Miller, 2006).

According to hazard management theory (Fiddick et al., 2000), a system well designed for reasoning about hazards and precautions should have properties different from one for detecting cheaters, many of which have been tested for and found (Fiddick, 1998, 2004; Fiddick et al., 2000; Pereyra & Nieto, 2004; Stone et al., 2002; Ermer et al., 2006). In addition to a specialization for reasoning about social exchange, we have therefore proposed that the human brain reliably develops computational machinery specialized for managing hazards, which causes good violation detection on precautionary rules. Obsessive-compulsive disorder, with its compulsive worrying, checking, and precaution taking, may be caused by a misfiring of this precautionary system (Boyer & Lienard, 2006; Cosmides & Tooby, 1999; Leckman & Mayes, 1998, 1999).

An alternative view is that reasoning about social contracts and precautionary rules is generated by a single mechanism. Some view both social contracts and precautions as deontic rules and wonder whether there is a general system for reasoning about deontic conditionals. More specifically, Cheng and Holyoak (1985, 1989) have proposed that inferences about both types of rule are generated by a permission schema, which operates over a larger class of rules, a class that encompasses social contracts and precautions.[14] Testing permission schema theory is particularly relevant to this chapter because it proposes production rules implementing inferences that might be found in a content-general deontic logic.

Can positing a permission schema explain the full set of relevant results? Or are they more parsimoniously explained by positing two separate adaptive specializations, one for social contracts and one for precautionary rules? We are looking for a model that is as simple as possible, but no simpler.

IV.v Social Contract Algorithms or a Permission Schema? Looking for Dissociations *within* the Class of Permission Rules (D1, D2, D4)

Permission rules are a species of conditional rule. According to Cheng and Holyoak (1985, 1989), these rules are imposed by an authority to achieve a social purpose, and they specify the conditions under which an individual is permitted to take an action. Cheng and Holyoak speculate that repeated encounters with such social rules cause domain-general learning mechanisms to induce a *permission schema*, consisting of four production rules (see table 2.3). This schema generates inferences about any conditional rule that fits the following template: "If action A is to be taken, then precondition R must be satisfied."

Social contracts fit this template. In social exchange, an agent permits you to take a benefit from him or her, conditional on your having met the agent's requirement. There are, however, many situations other than social exchange in which an action is permitted conditionally. Many customs and rules of etiquette are permission rules without being social contracts (e.g., "If one is to set the table with two forks, then the salad fork should

Table 2.3
The permission schema is composed of four production rules (Cheng & Holyoak, 1985)

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.

Social contracts and precautions fit the template of Rule 1:
Social contract: If the benefit is to be taken, then the requirement must be satisfied
Precaution: If the hazardous action is to be taken, then the precaution must be taken.

be on the outside"; "If you are to wear white shoes, you must do this between Memorial and Labor Days"). So are many of the baffling bureaucratic rules we are all subject to (e.g., "If you are turning in an intra-university memo, then it must be printed on blue paper"—we didn't make this one up!). Permission schema theory predicts uniformly high performance for the entire class of permission rules, a set that is larger, more general, and more inclusive than the set of all social contracts (see figure 2.5).

According to permission schema theory, a neurocognitive system specialized for reasoning about social exchange, with a subroutine for cheater detection, does not exist. Instead, it proposes that a permission schema causes good violation detection for all permission rules; social contracts are a subset of the class of permission rules; therefore, cheater detection occurs as a by-product of the more domain-general permission schema (Cheng & Holyoak, 1985, 1989). It is the closest the literature gets to proposing a content-general deontic logic machine.

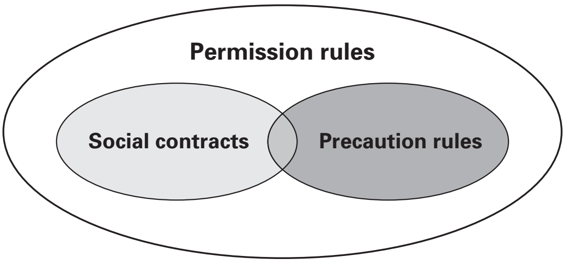


Figure 2.5
The class of permission rules is larger than, and includes, social contracts and precautionary rules. Many of the permission rules we encounter in everyday life are neither social contracts nor precautions (white area). Rules of civil society (etiquette, customs, traditions), bureaucratic rules, corporate rules—many of these are deontic conditionals that do not regulate access to a benefit or involve a danger. Permission schema theory (see table 2.3) predicts high performance for all permission rules; however, permission rules that fall into the white area do not elicit the high levels of performance that social contracts and precaution rules do. Neuropsychological and cognitive tests show that performance on social contracts dissociates from other permission rules (white area), from precautionary rules, and from the general class of deontic rules involving subjective utilities. These dissociations would be impossible if reasoning about all deontic conditionals were caused by a single schema that is general to the domain of permission rules (e.g., a "deontic logic machine"; see text).

In contrast, social contract theory holds that the design of the reasoning system that causes cheater detection is more precise and functionally specialized than the design of the permission schema. Social contract algorithms should have design features that are lacking from the permission schema, such as responsivity to benefits and intentionality. As a result, removing benefits (D1, D2) and/or intentionality (D4) from a social contract should produce a permission rule that fails to elicit good violation detection on the Wason task.

As Sherlock Holmes might put it, we are looking for the dog that did not bark: permission rules that do *not* elicit good violation detection. That discovery would falsify permission schema theory. Social contract theory predicts functional dissociations *within* the class of permission rules, whereas permission schema theory does not.

IV.vi No Benefits, No Social Exchange Reasoning: Testing D1 and D2

To trigger cheater detection (D2) and inference procedures specialized for interpreting social exchanges (D1), a rule needs to regulate access to benefits, not actions more generally. Does reasoning performance change when benefits are removed?

IV.vi.i Benefits Are Necessary for Cheater Detection (D1, D2) The function of a social exchange for each participant is to gain access to benefits that would otherwise be unavailable to them. Therefore, an important cue that a conditional rule is a social contract is the presence in it of a desired benefit under the control of an agent. *Taking a benefit* is a representational primitive within the social contract template *If you take benefit B, then you must satisfy requirement R*.

The permission schema template has representational primitives with a larger scope than that proposed for social contract algorithms. For example, *taking a benefit* is *taking an action*, but not all cases of taking actions are cases of taking benefits. As a result, all social contracts are permission rules, but not all permission rules are social contracts. Precautionary rules can also be construed as permission rules (although they need not be; see Fiddick et al., 2000, Experiment 2). They, too, have a more restricted scope: *Hazardous actions* are a subset of *actions; precautions* are a subset of *preconditions*.

Note, however, that there are permission rules that are neither social contracts nor precautionary rules (see figure 2.5). This is because there are actions an individual can take that are not *benefits* (social contract theory) and that are not *hazardous* (hazard management theory). Indeed, we encounter many rules like this in everyday life—bureaucratic rules, customs, and etiquette rules, for example, often state a procedure that is to be followed without specifying a benefit (or a danger). If the mind has a permission schema, then people should be good at detecting violations of rules that fall into the white area of figure 2.5, that is, permission rules that are neither social contracts nor precautionary. But they are not. Benefits are necessary for cheater detection.

Using the Wason task, several labs have tested permission rules that involve no benefit (and are not precautionary). As predicted by social contract theory, these do not elicit high levels of violation detection. For example, Cosmides and Tooby (1992) constructed Wason tasks in which the elders (authorities) had created laws governing the conditions under which adolescents are permitted to take certain actions. For all tasks, the law fit the template for a permission rule. The permission rules tested differed in just one respect: whether the action to be taken is a benefit or an unpleasant chore. The critical conditions compared performance on these two rules:

[15] "If one goes out at night, then one must tie a small piece of red volcanic rock around one's ankle."

[16] "If one takes out the garbage at night, then one must tie a small piece of red volcanic rock around one's ankle."

A cheater detection subroutine looks for benefits illicitly taken; without a benefit, it doesn't know what kind of violation to look for (D1, D2). When the permitted action was a benefit (getting to go out at night), 80% of subjects answered correctly; when it was an unpleasant chore (taking out the garbage), only 44% did so. This dramatic decrease in violation detection was predicted in advance by social contract theory. Moreover, it violates the central prediction of permission schema theory: that being a permission rule is sufficient to facilitate violation detection. There are now many experiments showing poor violation detection with permission rules that lack a benefit (e.g., Barrett, 1999; Cosmides, 1985, Experiments 5 and 6-C, plus all prescriptive clerical problems, referred to as "abstract" therein; Cosmides, 1989, Experiment 5; Fiddick, 2003; Griggs & Cox, 1982, Experiment 2; Griggs & Cox, 1983, Experiment 1; see discussion of deformed social contracts in Cosmides, 1985, pp. 60–64; Manktelow & Over, 1990, bingo problem; Platt & Griggs, 1993, Experiments 2 and 3).

This is another dissociation by content, but this time it is *within* the domain of permission rules—and, therefore, within the domain of deontic rules. To elicit cheater detection, a permission rule must be interpreted as restricting access *to a benefit*. It supports the psychological reality of the representational primitives posited by social contract theory, showing that the representations necessary to trigger differential reasoning are more content specific than those of the permission schema.

IV.vi.ii Benefits Trigger Social Contract Interpretations (D1) The Wason experiments just described tested D1 and D2 in tandem. But D1—the claim that benefits are necessary for permission rules to be *interpreted* as social contracts—receives support independent of experiments testing D2 from studies of moral reasoning. Fiddick (2004) asked subjects what justifies various permission rules and when an individual should be allowed to break them. The rules were closely matched for surface content, and context was used to vary their interpretation. The permission rule that lacked a benefit (a precautionary one) elicited different judgments from permission rules that restricted access to a benefit (the social contracts). Social agreement and morality, rather than facts, were more often cited as justifying the social contract rules. But facts (about poisons and antidotes), rather than social agreement, were seen as justifying the precautionary rule. Whereas most subjects thought it was acceptable to break the social contract rules if you were not a member of the group that created them, they thought the precautionary rule should always be followed by people everywhere. Moreover, the explicit exchange rule triggered very specific inferences about the conditions under which it could be broken: those who had received a benefit could be released from their obligation to reciprocate, *but only by those who had provided the benefit to them* (i.e., the obligation could not be voided by a group leader or by a consensus of the recipients themselves).

The inferences subjects made about the rules restricting access to a benefit follow directly from the grammar of social exchange laid out in social contract theory (Cosmides, 1985; Cosmides & Tooby, 1989). These inferences were not—and should not—be applied to precautionary rules (see also Fiddick et al., 2000). According to hazard management theory, the evolved function of mechanisms for reasoning about precautionary rules is to mitigate danger. This predicts precisely what Fiddick found: that facts about hazards and precautions are what justify precautionary rules in people's minds and make them important to follow. These inferences are expected if the deep representation of a precautionary rule is "If you are to engage in hazardous situation H, then you need to take precaution R *to increase your safety*." Words like "need," "must," or "ought" applied to the precaution do not specify an obligation to another person.

They specify a *necessary condition* for realizing a particular objective (increasing safety).

Social contracts and precautions may both be deontic, and in many cases both can be subsumed by a more general rule: "If condition C occurs, then a person *ought* to take action A." But the reasons behind that "ought" are vastly different for the two kinds of rules: safety for precautions, fulfilling an ethical obligation for social contracts. Moreover, the conditions under which this *ought* can (ethically) be ignored are different, so "ought" does not have the same moral implications in both cases. Lastly, the conditions that trigger the appropriate version of *ought* are different and content specialized. The *content* of condition C (*benefit* or *hazard*) and of action A (*provider's requirement* or *effective precaution*) need to be known to activate the domain-appropriate set of inferences.

Inferences about emotional reactions provide further evidence that social contracts and precautions activate two distinct inference systems rather than one permission schema. Fiddick (2004) asked subjects to predict which person (represented by faces with different emotional expressions) had seen someone violate a permission rule. He varied whether the permission rule provided contingent access to a benefit (i.e., was a social contract) or was precautionary. Social contract violations were thought to trigger anger, whereas precautionary violations were thought to trigger fear (Fiddick, 2004).

None of these dissociations within the realm of permission rules are predicted by permission schema theory. Moreover, because they involve different inferences and moral judgments for deontic rules drawn from different content domains, they would be difficult to explain on any content-free version of deontic logic. To be descriptively adequate, a deontic logic would have to include representational primitives like *hazardous activity*, *precaution*, *benefit*, and so on and trigger different sets of inferences when these are present. But a deontic logic with those properties would not be domain general. More likely, it would be a notational variant of the two different inferential systems proposed by social contract theory and hazard management theory.

IV.vii Intentional Violations versus Innocent Mistakes: Testing D4

Intentionality plays no role in permission schema theory. Whenever the action has been taken but the precondition has not been satisfied, the permission schema should register that a *violation* has occurred. As a result, people should be good at detecting violations of permission rules, whether the violations occurred by accident or by intention. In contrast, social contract theory predicts a mechanism that looks for *intentional* violations (D4).

Program designs that cause unconditional helping are not ESSs. Conditional helping can be an ESS because cheater detection provides a specific fitness advantage unavailable to unconditional helpers: by identifying cheaters, the conditional helper can avoid squandering costly cooperative efforts in the future on those who, by virtue of having an alternative program design, will not reciprocate. This means the evolutionary function of a cheater detection subroutine is to correctly connect an attributed disposition (to cheat) with a person (a cheater). It is not simply to recognize instances wherein an individual did not get what s/he was entitled to. That is, violations of social contracts are relevant for guiding future cooperative behavior insofar as they reveal individuals disposed to cheat—individuals who cheat by design, not by accident. Noncompliance caused by factors other than disposition, such as accidental violations and other innocent mistakes, does not reveal the disposition or design of the exchange partner. Accidents may result in someone's being cheated, but without indicating the presence of a cheater.[15] Indeed, program designs that refuse to cooperate with cooperators that may have made innocent mistakes do poorly in evolutionary simulations compared to those that forgive innocent mistakes and avoid cheaters (Panchanathan & Boyd, 2003).

Therefore, social contract theory predicts an additional level of cognitive specialization beyond looking for violations of a social contract. Accidental violations of social contracts will not fully engage the cheater detection subroutine; intentional violations will (D4).

IV.vii.i Accidents versus Intentions: A Dissociation for Social Contracts

Given the same social exchange rule, one can manipulate contextual factors to change the nature of the violation from intentional cheating to an innocent mistake. One experiment, for example, compared a condition in which the potential rule violator was inattentive but well-meaning to a condition in which she had an incentive to intentionally cheat. Varying intentionality caused a radical change in performance, from 68% correct in the intentional cheating condition to 27% correct in the innocent mistake condition (Cosmides, Barrett, & Tooby, forthcoming; supports D4). Fiddick (1998, 2004) found the same effect (as did Gigerenzer & Hug, 1992, using a different context manipulation).

In both scenarios, violating the rule would result in someone's being cheated, yet high performance occurred only when being cheated was caused by intentional cheating. Barrett (1999) conducted a series of parametric studies to find out whether the drop in performance in the innocent mistake condition was caused by the violator's lack of intentionality (D4) or by the violator's failure to benefit from her mistake (D2; see section IV.vi. on the necessity of *benefits* for eliciting cheater detection). He found that both factors independently contributed to the drop, equally and additively. Thus, the same decrease in performance occurred whether (i) violators would benefit from their innocent mistakes, or (ii) violators wanted to break the rule on purpose but would not benefit from doing so. For scenarios missing both factors (i.e., accidental violations that do not benefit the violator), performance dropped by twice as much as when just one factor was missing. That is, the more factors relevant to cheater detection were removed, the more performance dropped.

In bargaining games, experimental economists have found that subjects are twice as likely to punish defections (failures to reciprocate) when it is clear that the defector intended to cheat as when the defector is a novice who might have simply made a mistake (Hoffman, McCabe, & Smith, 1998). This provides interesting convergent evidence, using entirely different methods, for the claim that programs causing social exchange distinguish between mistakes and intentional cheating.

IV.vii.ii No Dissociation for Precautions Different results are expected for precautionary rules. Intentionality should not matter if the mechanisms that detect violations of precautionary rules were designed to look for people in danger. For example, a person who is not wearing a gas mask while working with toxic gases is in danger, whether that person forgot the gas mask at home (accidental violation) or left it home on purpose (intentional violation). That is, varying the intentionality of a violation should affect social exchange reasoning but not precautionary reasoning. Fiddick (1998, 2004) tested and confirmed this prediction: precautionary rules elicited high levels of violation detection whether the violations were accidental or intentional, but performance on social contracts was lower for accidental violations than for intentional ones. This functional distinction between precautionary and social exchange reasoning was predicted in advance based on the divergent adaptive functions proposed for these two systems.

That precautionary rules elicit no accident-intention dissociation, but social contracts do, is one more empirical fact about deontic reasoning that any domain-general deontic logic would need to somehow accommodate to be descriptively adequate. That a domain-general deontic logic would

W

distinguish intentional from accidental violations of deontic rules is certainly reasonable—even three-year-olds make such distinctions for social contracts (Núñez & Harris, 1998). It is less clear how a deontic logic could do this for some deontic rules and not others *without distinguishing rules by their content*.

A parallel dissociation between social exchange and precautions emerged in another series of experiments: We found that the moral character of the potential rule violator affects violation detection for social contracts but not precautionary rules (Cosmides, Tooby, Montaldi, & Thrall, 1999; Cosmides, Sell, Tooby, Thrall, & Montaldi, forthcoming). Subjects first read four scenarios in which "Mary" had opportunities to cheat in situations of social exchange. In one condition she cheated all four times; in the other she refrained from doing so. When Mary had been honest in these scenarios, this prior information relaxed cheater detection on the Wason task, but only for a social exchange in which Mary was the potential cheater (not for an identical task in which another person was the potential cheater). This is consistent with the notion that an honest person would only violate a social contract by mistake. In contrast, violation detection on a precautionary rule involving Mary was high whether the prior scenarios portrayed Mary as honest or dishonest in social exchange situations.

This dissociation between social exchange and precautionary reasoning has interesting implications for what kind of inferences about Mary's moral character subjects were extracting from the scenarios we provided. They were not extracting the inference that Mary is a deontic rule follower *in general*; if they had, the "Mary is honest" condition should have relaxed violation detection on the (deontic) precautionary rule. Instead, subjects were inferring that Mary can be trusted to fulfill her obligations in a specific type of situation: those involving social exchange.

IV.vii.iii Eliminating Permission Schema Theory The results of sections IV.vi and IV.vii violate central predictions of permission schema theory. According to that theory, first, all permission rules should elicit high levels of violation detection, whether the permitted action is a benefit or a chore, and second, all permission rules should elicit high levels of violation detection, whether the violation was committed intentionally or accidentally. Both predictions fail. Permission rules fail to elicit high levels of violation detection when the permitted action is neutral or unpleasant (yet not hazardous). Moreover, people are bad at detecting accidental violations of permission rules that are social contracts. Taken together, these results eliminate the hypothesis that the mind contains or develops a permission schema of the kind postulated by Cheng and Holyoak (1985, 1989).

IV.vii.iv Domain-General Deontic Logics Cannot Explain the Results It is sometimes proposed that cheater detection on social contracts is caused by the application of a domain-general deontic logic (for discussion of this possibility, see Manktelow & Over, 1987). The results presented in this section eliminate this hypothesis.

All the benefit and intentionality tests described in this section involved deontic rules, but not all elicited high levels of violation detection. This creates difficulties for any explanation that relies on the application of a deontic logic that is content blind. Permission schema theory attempted to achieve content generality by using deontic versions of *must* and *may* while employing very abstract representations of *actions* and *conditions*. This is similar to the move made in some deontic logics. But it *matters* whether the action to be taken is perceived as a benefit or a chore to the agent taking it: violation detection is poor when the action is a chore (IV.vi.i). Moreover, different moral inferences are made depending on whether the action is represented as a contingently provided *benefit* or a *hazard* (IV.vi.ii). Results showing that social contracts elicit good violation detection for intentional violations but not innocent mistakes (IV.vii.i), but that this dissociation does not exist for precautionary rules (IV.vii.ii) also pose a challenge to any deontic logic that does not distinguish rules by their domain. Importantly, the accidental-intentional dissociation poses a problem for any explanation that relies solely on how the rule is interpreted.

IV.vii.iv.i Interpretation Alone Is an Insufficient Explanation That social contract rules elicit good performance merely because we understand what implications follow from them (e.g., Almor & Sloman, 1996)—is eliminated by the intention versus accident dissociation. The same social contract rule, *with the same implications*, was used in both conditions. Access to a benefit (a good high school in the area) was conditional on a student's living in Dover City—a town where people pay high taxes to support this high school ("If a student is to be assigned to Dover High School, then that student must live in Dover City"). The story context explained the tax and school quality situation. Thus, the rule and its rationale were *identical* in all conditions. All that differed was the proposed *motivation* of the potential rule violators.

If the rule's implications were understood in the intention condition, they should also have been understood in the accident condition: it was the same rule with the same rationale. Yet the accident condition failed to elicit good violation detection. Understanding the implications of a social contract may be necessary for cheater detection, but the accident results show this is not sufficient.

Social contract theory contains a theory of interpretation (specific to situations involving social exchange), but it also posits a postinterpretive process: a subroutine that looks for cheaters. The accident-intention dissociation was predicted in advance of its discovery on the basis of the ESS-based task analysis of the adaptive problems posed by cheater detection, an analysis that predicted that the postinterpretive cheater detection mechanism would not only exist but have design features specialized for distinguishing accidental from intentional violations. To explain these results, social contract specific interpretive procedures *and* a postinterpretive cheater detection mechanism must work in tandem. (Both are also needed to explain results for switched rules and perspective change; Fiddick et al., 2000, contains an extended discussion of this issue in the context of relevance theory and its reliance on logical interpretation; Sperber et al., 1995.)

It is difficult to see how the motivations of potential violators could play a role in any explanation based solely on interpretation of the implications of a conditional rule. But let's assume for the sake of argument that a deontic logic was constructed that could somehow explain why the accident/intention distinction matters for reasoning about social contracts. To account for the empirical facts, this explanation would also have to explain why the violator's motives matter for social contracts but *not* for precautionary rules. Remember that precautionary rules do not elicit a dissociation in violation detection based on accident versus intention. Any explanation of this difference between precautionary and social contract rules would have to refer to the fact that their *content* is different. And a deontic logic that distinguished the two by their content would no longer be content-free.

IV.vii.iv.ii Eliminating Fodor's Artifact Hypothesis The results of sections IV.vi and IV.vii also defeat a related claim by Fodor (2000): that "the putative cheater detection effect on the Wason task is actually a materials artifact" (p. 29). This sweeping conclusion is predicated on the (mistaken) notion that the only evidence for cheater detection comes from experiments in which the control problems are indicative (i.e., descriptive) conditionals (a curious mistake because it is refuted by experiments with *deontic* controls, which are presented in the single source Fodor cites: Cosmides & Tooby, 1992). According to Fodor (2000), the fact that people are good at detecting violations of social contracts but not indicative rules "is built into a difference between *the logic* of indicative and deontic conditionals" (emphasis added). He argues that deontic and indicative conditionals are "really about" different things and presents an argument about what "the" correct parsing (the correct interpretation) of each rule type "is." (Forgive us for seeing naive realism in this language—or, more charitably, the kind of loose talk that breeds ontological confusion.)[16]

Fodor's explanation strikes us as deeply flawed—among other things, it assumes what it seeks to explain (which he more or less acknowledges in his footnote 5). He argues that, whereas indicative conditionals are really about *P* implying *Q*, deontic conditionals are really about mandating *Q*[17] and are therefore correctly parsed as *Required: Q (in the case that P)*. For example, the car rule would be assigned the interpretation *Required: fill the tank with gas (in the case that you borrow the car)*. Because subjects can reason with the law of noncontradiction, he argues, they choose the *not-Q* card when asked to look for cases in which *required: Q* was violated. They also know to check whether the condition that triggers the requirement holds—that is, whether it is the case that *P*—so they choose the *P* card.

This account is problematic for two reasons. First, the reason given for choosing the cards is not really an explanation: it is just a redescription of the correct answer (which subjects give in response to standard social contracts). Second, it is not clear why Fodor's explanation shouldn't also apply to indicative conditionals. As Quine (1972, p. 19) points out, some logicians interpret indicative conditionals as meaning *necessarily: Q (in the case that P)*. This entirely reasonable interpretation is parallel to the logical form Fodor proposes for deontic conditionals. On this view, when given an indicative conditional and asked to look for violations, subjects will look for cases that violate *necessarily: Q*, thereby choosing the *not-Q* card (because they can reason with the law of noncontradiction). They should also choose the *P* card to check that the condition triggering the necessary presence of *Q* holds. That is, indicative conditionals should elicit high levels of "*P* & *not-Q*." Yet they do not.

In another passage, Fodor rejects the notion that cheater detection involves the concept of obligation (in favor of *requirement*) on the basis of (a mangled version of) the drinking age law (the high-performing version of which fits the social contract template). He says this rule cannot *obligate* one to be over 21 years old because one cannot be obligated to be other

W

than one is.[18] Perhaps so. But the drinking age law does not obligate anyone to be 21 years old; it obligates people to *wait* until they are 21 years old before drinking beer. In social exchange, the function of providing *contingent* access to a benefit (like drinking beer) is to create a situation that benefits the agent who controls access to that benefit. In creating the drinking age law, an agent (a social group) obligates people to wait until they are old enough to behave in a responsible manner before granting access to the benefit (beer).[19] *This creates a situation that benefits the social group that made the rule*: by preventing teen-agers from drinking, it cuts down on drunk driving accidents and other negative externalities caused by drunken youths.[20] More generally, the requirement in social exchange is imposed to create a *situation* that benefits the provider. If it benefits the provider to require that a person have certain properties in order to be eligible to receive the benefit the provider is offering, so be it.

These are the problems we have with Fodor's reasoning. But instead of focusing on these issues, let us consider whether his artifact explanation can account for the cheater detection results observed. After all, there are many experiments comparing reasoning on social contracts to reasoning about other *deontic* conditionals—ones which should be assigned precisely the same interpretation *Required: Q (in the case that P)*.

According to Fodor, high levels of violation detection will be found for any deontic rule that specifies what people are (conditionally) required to do because *not-Q* responses will be elicited by *Required: Q (in the case that P)*. All the permission rules described in section IV.vi.i had precisely this property, all were stipulated to be rules that are in effect and, in every case, subjects were asked to reason from the rule, not about it. For example, by Fodor's account, rules [15] and [16] have precisely the same logical form: *Required: that one tie a small piece of red volcanic rock around one's ankle (in the case that P)*. They differ *only in what P refers to*. In [15] it is something adolescents see as a benefit—going out at night—and in [16] it is something they see as a chore. But what *P* refers to plays no role in Fodor's account. If Fodor's artifact hypothesis were correct, both [15] and [16] should have elicited good violation detection. But [16] did not. Violation detection was poor when the deontic rule lacked a benefit, in this experiment as well as in all the others cited in section IV.vi.i.

The failure of Fodor's account is even more apparent in the accident-intention experiments (IV.vii.i). The social contract rule used was identical in all conditions, so, by Fodor's account, it would be assigned the same logical form in all conditions. Yet performance tracked the motivations of the violators: there was poor violation detection in the innocent mistake condition but good violation detection in the intentional cheating condition. Motives to violate can play no role in a "materials artifact" explanation like Fodor's. (Buller, 2005, adopted Fodor's explanation in his chapter on social exchange reasoning, similarly ignoring the data on dissociations within the domain of deontic rules. Thus, his analysis suffers from the same inability to account for the facts.)

Wason tasks involving social contracts have been compared to ones involving other types of deontic conditionals since our very first experiments in the early 1980s. The evidentiary basis for the existence of social contract algorithms, equipped with a cheater detection mechanism, has always included dissociations in performance *within* the domain of deontic rules.

IV.vii.v Implications for Moral Reasoning? The results of sections IV.vi and IV.vii show that it is not enough to admit that moral reasoning, social reasoning, or deontic reasoning is special. The computational machinery engaged when people reason about social exchange shows a specificity of design that is far narrower in scope. Deontic rules expressing social contracts elicit different patterns of reasoning than other deontic rules, whether precautionary or otherwise. Moreover, reasoning about social exchange is regulated by factors that have no impact on reasoning about other deontic rules, including information about the potential violator's intentions and moral character as a cooperator.

In the next section we will see that reasoning about social contracts and precautionary rules is not only functionally distinct, but it is associated with different areas of the brain.

IV.viii A Neuropsychological Dissociation between Social Contracts and Precautions

Like social contracts, precautionary rules are conditional, are deontic, and involve subjective utilities. Moreover, people are as good at detecting violators of precautionary rules as they are at detecting cheaters on social contracts. This has led some to conclude that reasoning about social contracts and precautions is caused by a single more general mechanism (e.g., general to permissions, to deontic rules, or to deontic rules involving subjective utilities; Cheng & Holyoak, 1989; Manktelow & Over, 1988, 1990, 1991; Sperber et al., 1995). Most of these one-mechanism theories are undermined by the series of very precise, functional dissociations between social exchange reasoning and reasoning about other deontic permission rules discussed above. However, a very strong test, one that addresses *all* one-mechanism theories, would be to find a neural dissociation between social exchange and precautionary reasoning.

IV.viii.i One Mechanism or Two? If reasoning about social contracts and precautions is caused by a single mechanism, then neurological damage to that mechanism should lower performance on both types of rule. But if reasoning about these two domains is caused by two functionally distinct mechanisms, then it is possible for social contract algorithms to be damaged while leaving precautionary mechanisms unimpaired, and vice versa.

Stone et al. (2002) developed a battery of Wason tasks that tested social contracts, precautionary rules, and descriptive rules. The social contracts and precautionary rules elicited equally high levels of violation detection from normal subjects (who scored 70% and 71% correct, respectively). For each subject, a difference score was calculated: percentage correct for precautions minus percentage correct for social contracts. For normal subjects, these difference scores were all close to zero ($M = 1.2$ percentage points, $SD = 11.5$).

This battery of Wason tasks was administered to R.M., a patient with bilateral damage to his medial orbitofrontal cortex and anterior temporal cortex, plus damage near the posterior temporal poles sufficient to disconnect both of his amygdalae. R.M.'s performance on the precaution problems was 70% correct: equivalent to that of the normal controls. In contrast, his performance on the social contract problems was only 39% correct. R.M.'s difference score (precautions minus social contracts) was 31 percentage points. This is 2.7 standard deviations larger than the average difference score of 1.2 percentage points found for control subjects ($p < .005$). In other words, R.M. had a large deficit in his social contract reasoning, alongside normal reasoning about precautionary rules.

Double dissociations are helpful in ruling out differences in task difficulty as a counterexplanation for a given dissociation (Shallice, 1988), but here the tasks were perfectly matched for difficulty. The social contracts and precautionary rules given to R.M. were logically identical, posed identical task demands, and were equally difficult for normal subjects. Moreover, because the performance of the normal controls was not at ceiling, ceiling effects could not be masking real differences in the difficulty of the two sets of problems. In this case, a single dissociation licenses inferences about the underlying mental structures. R.M.'s dissociation supports the hypothesis that reasoning about social exchange is caused by a different computational system than reasoning about precautionary rules: a two-mechanism account.

IV.viii.ii Neuroimaging and Rule Interpretation Recent functional magnetic resonance imaging studies also support the hypothesis that social contract reasoning is supported by different brain areas than precautionary reasoning (Wegener, Baare, Hede, Ramsoy, & Lund, 2004; Fiddick, Spampinato, & Grafman, 2005). We recently conducted a neuroimaging study comparing reasoning on Wason tasks involving social contracts to ones involving (i) precautionary rules and (ii) indicative rules involving social behavior (Ermer et al., 2006). Like the other studies, we found that reasoning about social exchange activates brain areas not activated by reasoning about precautionary rules, and vice versa. Unlike the other studies, however, the design of the Ermer et al. study allows one to distinguish brain activations while the rule is being read and interpreted from brain activations at the postinterpretive phase, when subjects are deciding whether a card should be turned over to detect violations. Social contracts and precautions activated different brain areas during both stages, supporting the hypothesis that the postinterpretive detection process *and* the interpretive process differ for these two, content-defined classes of deontic conditional. The results for the interpretation phase were particularly illuminating.

Baron-Cohen (1995) proposed that the theory-of-mind inference system evolved to promote strategic social interaction. Social exchange—a form of cooperation for mutual benefit—involves strategic social interaction (remember the Prisoners' Dilemma) and requires theory-of-mind inferences about the contents of other individuals' mental states, especially their *desires*, *goals*, and *intentions*. Indeed, inferences about the goals and desires of agents are necessary for situation detectors to recognize an interaction as involving social exchange (see section II.i.i). Thus, one might expect neural correlates of theory of mind to be activated when subjects are interpreting social exchange rules. That is precisely what Ermer et al. found. Anterior and posterior temporal cortex—both previously identified in the literature as neural correlates of theory-of-mind inferences—were activated when subjects interpreted social exchange rules, but not when they were interpreting precautionary rules.

Figure 2.6 shows the average signal intensity in the anterior temporal cortex for each individual social contract and precautionary problem (anterior temporal cortex was extensively damaged in R.M.). The figure shows that there is almost no overlap for these two sets of rules: signal intensities do not overlap at all for 14 of 16 problems (7 of 8 social contracts and 7 of 8 precautions). That is, the greater activation of anterior temporal cortex for social exchange rules compared to precautionary rules is systematic—it is not an artifact of one or two problems. It also shows the results are not

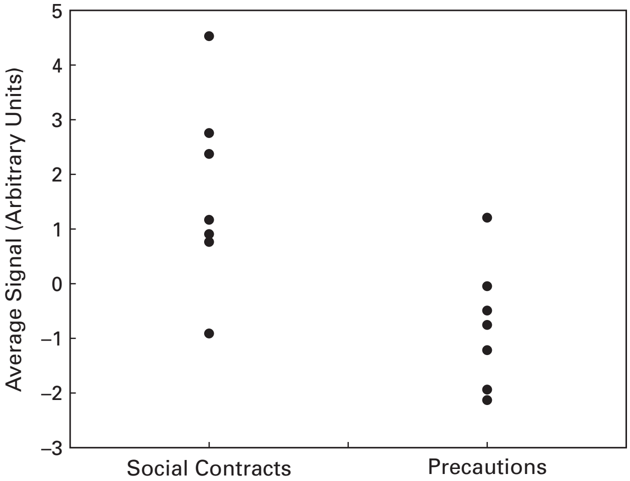


Figure 2.6
Average signal intensity in anterior temporal cortex when reasoning about social contracts and precautionary rules. Each point represents an individual reasoning problem. There is virtually no overlap in signal intensity between the social contract and precautionary rules. That this pattern of differential activation replicates across *individual problems* is expected if the activation differences reflect the underlying, content-specific representation and interpretation of social exchange versus precautionary problems.

caused by idiosyncratic, theory-irrelevant properties of individual rules. The surface content of the various social contract rules used—their particular antecedents and consequents—differed across problems; so did the surface content of the various precautionary rules used. The social exchange problems were similar to one another only by virtue of fitting the situation of social exchange described in section II.i.i and the benefit-requirement template of a standard social contract. Likewise, the precautionary rules were similar to one another only by virtue of fitting the hazard-precaution template for a precautionary rule specified in section IV.iv.i. That the pattern of differential activation replicates across *individual problems* increases our confidence that these activation differences reflect the underlying, content-specific representation and interpretation of social exchange versus precautionary problems.

These results are consistent with our task analyses of the two domains. Inferences about the content of other people's mental states—theory-of-mind inferences—are necessary for interpreting rules involving social exchange but not for interpreting precautionary rules. To interpret a rule as precautionary requires the ability to recognize facts about the world: that an activity may be hazardous and that taking a particular precaution may mitigate that hazard (sections IV.iv.i, IV.vii.ii, IV.vii.ii). No inferences about mental states are required. This result adds to the evidence that the mind distinguishes deontic rules by their content and that the interpretive process applied to social exchange rules is different from that applied to precautionary rules.

IV.viii.iii Eliminating One-Mechanism Hypotheses Every alternative explanation of cheater detection proposed so far claims that reasoning about social contracts and precautions is caused by the same neurocognitive system. R.M.'s dissociation is inconsistent with all of these one-mechanism accounts. These accounts include mental logic (Rips, 1994), mental models (Johnson-Laird & Byrne, 1991), decision theory/optimal data selection (Kirby, 1994; Oaksford & Chater, 1994), permission schema theory (Cheng & Holyoak, 1989), Fodor's deontic logic account (Fodor, 2000; Buller, 2005), relevance theory (Sperber et al., 1995), and Manktelow and Over's (1991, 1995) view implicating a system that is general to any deontic rule that involves subjective utilities. (For further evidence against relevance theory, see Fiddick et al., 2000; for further evidence against Manktelow & Over's theory, see Fiddick & Rutherford, 2006.)

Indeed, no other reasoning theory even distinguishes between precautions and social contract rules. The distinction is derived from evolutionary-functional analyses and is purely in terms of *content*. These results, together with the others discussed in section IV, indicate the presence of a very narrow, content-sensitive cognitive specialization within the human reasoning system for reasoning about social exchange.

V Conclusion

Philosophers have already expressed the wish to construct a deontic logic that captures important facts about human deontic reasoning. A fact of signal importance for the success of this project emerges from the research we have discussed: Deontic reasoning is not a unified phenomenon. Reasoning about deontic conditionals fractionates in a way implying the existence of at least two specialized systems: one for reasoning about social exchange and another for reasoning about precautionary rules. There may be others as well (Cosmides & Tooby, 2006; Tooby, Cosmides, & Price, 2006).

Deontic conditionals expressing social exchanges activate reasoning machinery that is very precisely engineered for producing an evolutionarily stable form of cooperation. When situation detectors register the presence of cues that fit the input conditions for social exchange (cues indicating what two agents want and their willingness and ability to provide these benefits to one another), they activate a domain-specialized computational system, the *social contract algorithms*. Social contract algorithms import surplus structure as they continue to interpret the situation, including any deontic conditional that has been stated, inserting very specific deontic concepts of *obligation* and *entitlement* in just the right places. Words such as "must," "may," and "ought," which can refer to very different concepts depending on context, are thus assigned an exchange-appropriate meaning. Indeed, interpretive procedures within the social contract algorithms ensure that concepts of obligation and entitlement are understood to apply, even when deontic words are missing entirely.

The social contract algorithms include a domain-appropriate set of inferential rules, which are applied spontaneously and intuitively to situations of social exchange. These license inferential transformations that are "common sense" when applied to social exchange but that violate our moral intuitions when applied to other deontic domains. They also generate very predictable moral judgments about the circumstances under which the obligations of a social contract can be lifted or ignored without blame—circumstances that do not apply to precautionary and other deontic conditionals. They include a specialized information search procedure as well—the cheater detection mechanism—that looks for information that could reveal potential cheaters, very narrowly defined. By seeing what conditions turn the cheater detection mechanism on and off, we found that its procedures define a cheater as an individual who has (i) taken the benefit the provider agreed to supply contingent on a requirement being satisfied, (ii) done so without having satisfied the provider's requirement, and (iii) taken these actions by intention rather than by accident.

Supporting the claim that reasoning about social exchange is functionally distinct from reasoning about other deontic conditionals, brain damage can impair social exchange reasoning selectively, leaving intact one's ability to reason well about deontic precautionary rules. Neuroimaging results are consistent with this finding: different patterns of brain activation are produced by reasoning about social contracts versus precautionary rules, during interpretation *and* during the violation detection process. Significantly, neural correlates of theory-of-mind are differentially activated when people interpret social exchanges compared to precautionary rules.

The behavioral data show that deontic conditionals expressing precautions activate interpretive procedures, deontic concepts, moral judgments, and violation detection procedures that differ from those activated in social exchange. Moreover, deontic conditionals that tap neither domain do not elicit the enhanced reasoning competence found for social exchange and precautionary rules. Parametric studies (Barrett, 1999; Cosmides, Barrett, & Tooby, forthcoming) and priming results (Cosmides, 1985; Cosmides & Tooby, 1987; Fiddick, 1998) suggest that those people who do reason correctly about such rules succeed when the rules contained enough input cues consistent with social exchange or precautions to weakly activate one of these specialized systems.

We should not imagine that there is a separate specialization for solving each and every adaptive problem involving deontic reasoning (Tooby, Cosmides, & Price, 2006). Nor should real differences in processing be ignored in a misguided effort to explain all reasoning performance by reference to a single mechanism. As Einstein once said, "Make everything as simple as possible, but no simpler." We think the same principle should apply to creating a deontic logic that is descriptively accurate.

A deontic calculus that is content blind would have difficulties capturing the many distinctions our minds spontaneously make between domains of deontic reasoning—we have tried to show this for each result presented. The data suggest the need for more than one deontic logic, each associated with domain-specialized inferential rules and content-defined scope boundaries (ones that respect the domain boundaries of evolved reasoning specializations); each calculus may also need some domain-specialized deontic operators to capture evolved distinctions between deontic concepts (especially insofar as one views the meaning of concepts as defined by the system of inferences in which they participate). This does not mean, however, that deontic logicians will have to create an endless series of formal calculi. There may be a kaleidoscopic array of culturally defined deontic domains (with different sets found in different cultures), but it is likely that these arose through the application of a much smaller set of evolved social inference systems (e.g., Fiske, 1991).

As Sperber (1994, 1996) has argued, an evolved mechanism's actual domain (the entire set of circumstances in which it is applied) is larger than its proper domain (the set of circumstances that selected for its evolved design). This is because an evolved system is a causal system: it will be activated whenever there are cues that fit the mechanism's input conditions. For example, other humans—agents with desires, goals, intentions, emotions, and beliefs—are the proper domain of the social contract algorithms. But people in many places are told that unseen agents—gods, ghosts, and spirits—interfere with events in pursuit of their own goals and desires. Social inference machinery is activated by these representations of agents and their desires, with predictable consequences: people propose social exchanges to their gods, promising ritual offerings, sacrifices, or reform in exchange for help (Boyer, 2001). Similarly, precautionary inferences—including ones relevant to contagion and disgust—are often applied in religious ceremonies (especially those involving dead bodies), and in certain caste systems, a path-dependent history of interaction with corpses and other polluting substances has led to the inference that certain categories of people are themselves polluted (Boyer, 2001; Fiddick, 1998, 2004; Haidt & Joseph, 2004). Failing to take the appropriate precautions when interacting with members of another caste is often considered a moral transgression, although presumably of a different kind than cheating a friend or murdering a neighbor. In the realm of political ideas, representing a citizen's relationship to government as a form of social exchange triggers the inference that paying taxes is fulfilling an obligation rather than surrendering to extortion (for more examples, see Cosmides & Tooby, 2006). That is, a small set of evolved social inference systems may give rise to a diverse array of cultural forms. In formulating domain-specific deontic logics, it may be more parsimonious to first consider the boundaries respected by the evolved architecture of the mind. As a later exercise, logicians and moral philosophers could consider metarules for adjudicating cases in which it is unclear which formal calculus should be applied.

At this point, some readers are surely thinking that the goal of capturing facts about deontic reasoning should be abandoned so as to preserve the hope of creating a single, simple deontic logic that is truly domain general. This would be unwise. First, there is the problem we raised at the beginning: the domain specificity of our deontic intuitions is likely to still be present in what is supposed to be a general deontic logic, but disguised as different kinds of "reasons" (or swept under the rug in other ways). A second problem arises when one considers the intersection of deontic logic, moral philosophy, and the epidemiology of cultural ideas, including moral ones.

By applying what is known about the evolved architecture of the human mind to various problems in cultural transmission, cognitive anthropologists are starting to understand why some ideas are contagious, spreading easily from mind to mind, whereas others are proposed and soon forgotten (Boyer, 2001; Sperber, 1996). Ideas that violate evolved intuitions can be attention grabbing (e.g., "an electron is a particle that behaves like a wave!"), but those that do not simultaneously activate an evolved system that supports rich inferences (e.g., the object mechanics inference system; Spelke, 1990; Leslie, 1994) remain the preoccupation of specialists (e.g., quantum physicists). A deontic calculus that violates too many evolved moral intuitions is likely to have a similar fate. This should concern any moral philosopher for whom outcomes matter.

Whereas some philosophers argue that an outcome is ethical if the procedure that produced it was ethical (e.g., Nozick, 1974), others are consequentialists: they argue that certain outcomes are ethically better than others and that moral choices should be based—at least in part—on how well they achieve ethical outcomes (e.g., Bentham, 1789; Rawls, 1971; Sen, 1989). Consequentialists need to be concerned with consequences: their job is not merely to define what end state is morally preferable but to elucidate methods that are likely to achieve that end state. Moral action depends, at least in part, on moral reasoning. Versions of deontic logic that capture empirical facts about deontic reasoning are likely to be intuitively compelling, easy to understand, yet precise enough to clarify moral questions in a way that can promote ethical choices. Versions that neglect these facts are less likely to be understood, accepted, or generally adopted. A consequentialist should not prefer a deontic logic merely because it is domain general. A consequentialist needs to consider whether a given deontic logic is likely to be widely adopted and used to inform the multitude of real-world decisions that shape our social world. When outcomes matter, human nature matters.

Notes

We warmly thank Walter Sinnott-Armstrons, whose insights and guidance made this chapter possible.

1. Even Castañeda (1981), whose approach sometimes distinguishes contexts, implicitly endorses this domain-general hope when he says, "a formal calculus proposed as a deontic calculus together with *its* primary interpretation is a theory about the logical structure of our ordinary deontic language and about our ordinary deontic reasonings" (38, emphasis added). "Its primary interpretation" implies a single primary interpretation, not a multiplicity of domain-specific ones. We thank Walter Sinnott-Armstrong for pointing this out.

2. A later project might be to develop a metatheory for deciding which deontic logic should apply in situations that (arguably) fall into two or more domains.

3. In this early work we referred to deontic rules as "prescriptive" (because they prescribe behavior) and indicative ones as "descriptive" (because they describe behavior or facts about the world).

4. This is a regularity in present environments as well, but the inference is obviously not induced through ontogenetic experience. First, mental states cannot be seen;

one can observe correlations between behavior and eye direction, but not between mental states and eye direction. Second, people with autism are quite capable of noticing correlations in the world, yet they do not induce this one.

5. A system can be ecologically rational yet operate over a very broad array of content. Examples would be the frequency computation system, the fast-and-frugal heuristics identified by Gigerenzer and colleagues, and the system underpinning classical conditioning, which Gallistel describes as implementing nonstationary, multivariate time-series analysis (Brase, Cosmides, & Tooby, 1998; Cosmides & Tooby, 1996a; Gigerenzer et al., 1999).

6. Surreally, this paper, which highlights interpretation in its title, is cited by Buller (2005), in the same chapter in which he claims that we do not realize that different conditional rules "have" different interpretations.

7. If the rules regulating reasoning and decision making about social exchange do not implement an ESS, it would imply that these rules are a by-product of some other adaptation that produces fitness benefits so huge that they compensate for the systematic fitness costs that result from its producing non-ESS forms of social exchange as a side effect. Given how much social exchange humans engage in, this alternative seems unlikely.

8. The same logic applies even if no cost is incurred. A design that provides benefits contingent on a benefit being provided in return will have better fitness than one that provides benefits unconditionally (Tooby & Cosmides, 1996).

9. What that "more" consists of for indicative rules is a matter of debate. Sperber, Cara, and Girotto (1995) argue that the conditional rule must first be interpreted as pragmatically implying a denial (*deny: P and not-Q*). They have a similar argument for social contracts, arguing that these are interpreted as forbidding (*forbid: P and not-Q*). Whether their argument is correct for indicatives, their extension to all deontic rules (including social contracts) is not. As we discuss in Fiddick et al. (2000), their theory does not predict that conditionals expressing trades will be interpreted as meaning *forbid: P & not-Q* (nor will most forms of social exchange). Therein we further show that their content-general interpretive procedures (employing logical equivalences) cannot explain results for switched social contracts, perspective change, or, indeed, any of the dissociations within the domain of deontic rules discussed in section IV.

10. *Programs that cheat by design* is a more general formulation of the principle, which does not require the human ability to form mental representations of intentions or to infer the presence of intentional mental states in others. An analogy to deception may be useful: birds that feign a broken wing to lure predators away from their nests are equipped with programs that are designed to deceive the predator, but the cognitive procedures involved need not include a mental representation of an *intention* to deceive.

11. First-order logic cannot solve the problem of cheater detection even if one assumes the existence of social contract algorithms that are interpreting rules and importing surplus structure into them—see Fiddick, Cosmides, and Tooby (2000) for an extended discussion of this point.

12. And one who has not filled the tank and not borrowed the car has merely decided not to take the parents up on their offer.

13. Moreover, first-order logic contains no rules of inference that allow *If P, then Q* to be translated as *If Q, then P* (i.e., no rule for translating [1] as [2]; see text) and then applying the logical definition of violation to that translation (for discussion, see Fiddick et al., 2000).

14. Cheng and Holyoak (1985) also propose an obligation schema, but permission and obligation schemas do not lead to different predictions on the kinds of rules usually tested (for discussion, see Cosmides, 1989; Rips, 1994, p. 413).

15. Mistakes can be faked, of course. Too many by a given individual should raise suspicion, as should a single mistake that results in a very large benefit. Although this prediction has not been tested yet, we would expect social contract algorithms to be sensitive to these conditions.

16. For example, is Fodor talking about properties of logic, independent of the structure of human minds? Or is he talking about properties of human minds as they interpret utterances? If it is the latter, then he is agreeing with us that the mind interprets social contracts differently from indicatives, but without providing an explanation for why this occurs. (Presumably he is not claiming that ordinary intuition is carrying out the reductio argument that he himself had difficulty constructing; see his footnote 6.)

17. Fodor doesn't appear to understand what we mean by a "social contract" (indeed the term does not appear in his piece). When I say, "If you give me your watch, then I'll give you $100" it is not to prohibit *not-Q* or to mandate *Q*: it is to propose a trade. For an extensive analysis of the problems with interpreting social contracts as primarily about prohibition, see Fiddick, Cosmides, and Tooby (2000, especially pp. 35–41).

18. Indeed, this "counterexample" to an implication of his reductio is what leads him to reject the proposition that *It is required that: If someone is under 18, s/he drinks coke* "isn't really about P → Q being required" (Fodor 2000, p. 31); that it is instead about *Q* being required (in the case that *P*).

19. Indeed, changing conceptions of what counts as "old enough to behave responsibly" are why the age was raised from 18 to 21.

20. To the extent that a subject interprets "drinking beer" as a hazardous activity, the drinking age law could be interpreted as a precautionary rule; because it has a dual interpretation, it falls into the area of intersection between social contracts and precautions on figure 2.5.