Deduction or Darwinian Algorithms?

An explanation of the "elusive" content effect on the Wason selection task

A thesis presented

by

Leda Cosmides

to

The Department of Psychology and Social Relations

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Psychology

Harvard University

Cambridge, Massachusetts

July, 1985

© 1985 by Leda Cosmides

All rights reserved.

Abstract

This thesis develops the idea that natural selection has shaped how humans reason about evolutionarily important domains of human activity. The human mind can be expected to include "Darwinian algorithms" that are specialized for processing information about such domains. Evolutionary principles were heuristically applied to pinpoint social exchange as an adaptively important domain of human activity; these principles were also applied in developing computational theories of how humans process information about social exchange. Evidence is presented supporting the hypothesis that the human mind includes Darwinian algorithms specialized for reasoning about social exchange. This hypothesis both predicts and explains "content effects" on the Wason selection task -- a test of logical reasoning -- better than alternative theories.

Table of Contents

Deduction or Darwinian Algorithms?
An explanation of the "elusive" content effect
on the Wason selection task

Chapter 1
Logic and the Study of Human Reasoning

Chapter 2
A review of the literature on the "elusive" content effect
on the Wason selection task

Chapter 3
"Differences in Experience":
Proposed explanations for the elusivity of the
content effect on the Wason selection task

Chapter 4

Darwinian Algorithms

Chapter 5

Human Social Exchange

Chapter 6

Social contracts and the Wason selection task: Experiments

Chapter 7

Discussion and Conclusions

Introduction

The equipotentiality assumption has crept, incognito, from the meta-theory of behaviorism* into the meta-theory of cognitive psychology. Behaviorists do not expect the laws of learning to differ from domain to domain; cognitive psychologists do not expect the processes that govern attention, memory, or reasoning to differ from domain to domain.

To the behaviorist, stimuli are stimuli and responses, responses: their content is not supposed to affect how they are paired. When content effects are discovered, the behaviorist speaks of adjusting "parameter values", or of differences in the organism's "experience" with various content domains. To the cognitive psychologist, information is information: the content of the information is not supposed to affect how it is processed. When content effects are discovered, the cognitive psychologist also speaks of differences in the organism's "experience" with various content domains.

Though unspoken, the message is clear: Content is noise. Cognitive processes are content-independent, domain general, equipotential. The human mind is a general purpose information processing system, designed to process any kind of information with equal efficiency. The format of the information -- for example, whether it is imagistic or propositional -- might make a difference in how it is processed, but its content will not. The amount of experience the organism has had with a domain may affect performance, but correct for this and the effect will

---

* for review, see Herrnstein, 1977.

disappear. These claims are rarely tested; they are merely assumed.

The alternative view -- that the human mind includes a number of domain specific, content-dependent, information processing systems -- is seldom entertained (Cf. Chomsky, 1975; Marr & Nishihara, 1978; Fodor, 1983). Although behaviorism came under brisk attack from evolutionary and ethological quarters for assuming that learning was equipotential (Herrnstein, 1977), cognitive psychology seems untouched by this fray and the serious problems it raised. Yet the evolutionary arguments against the equipotentiality assumption in behaviorism apply equally to cognitive psychology.

An unspoken assumption is an unexamined assumption. Cognitive processes may, in fact, be content-independent; if so, then this should be proved, not presumed. Indeed, when content effects are found, the content-independence of cognitive processes should be a hypothesis of last resort.

Cognitive processes, like electrons, are entities defined solely by input-output relations. An electron gun is fired at a diffraction slit, and then into a cloud chamber: even though the data from the first firing indicates a wave and the data from the second indicates a particle, there are compelling reasons for believing these divergent patterns were created by one and the same entity. It would grossly violate our most basic notions of similarity and causation to categorize two firings of an electron gun as two different "stimuli", just because they were fired at different targets. The same input -- which was, in this case, the very entity physicists were trying to characterize -- yielded

different outputs. The only reasonable theoretical alternative was to complexify the equations defining the electron, and assume that it did not correspond to any ordinary human concept like "particle" or "wave" (Heisenberg, 1971).

But there are no compelling reasons -- other than a misguided sense of parsimony -- for believing that the same cognitive process is involved when manifestly different inputs yield unmistakably different outputs. In fact, because cognitive processes are entities defined by these very input-output relations, the discovery of content effects should be taken as prima facie evidence that the different stimuli tested are accessing different cognitive processes. If response patterns vary with stimulus content, but their variation does not appear to be systematic, then one should rethink one's theory of how to parse the world into content domains. Hand-waving appeals to "differences in experience" -- which are virtually impossible to falsify -- should be explanations of last resort.

When content effects are found, cognitive psychologists should entertain the hypothesis that domain specific, content-dependent, cognitive processes are responsible. Content effects have been found on the Wason selection task, a famous experimental paradigm that tests whether people reason according to the content-independent canons of formal logic. Most attempts at explaining these content effects have appealed to "differences in subjects' experience" with various content domains. A controversy has grown up around these content effects, because, to date, they have eluded prediction.

This thesis uses content effects on the Wason selection task

to test the hypothesis that humans have domain specific, innate mental algorithms specialized for reasoning about social exchange. A computational theory of the functional properties of these algorithms was derived using natural selection theory as a heuristic guide. Critical tests were conducted to choose between the social exchange hypothesis and the hypotheses in the literature that appeal to "differences in experience."

The discovery of systematic variation from domain to domain is strong evidence that domain specific algorithms are at work; so is the discovery of systematic variation within a domain that cannot be easily explained by a content-independent process. Both kinds of evidence are presented in support of the social exchange hypothesis. I argue that no other hypothesis offered so far can predict or explain the experimental results presented herein, and that the social exchange hypothesis best explains the content effects on the Wason selection task that have already been reported in the literature.

The meta-theoretical view entailed by this hypothesis -- of the human mind as a collection of functionally distinct, Chomskian "mental organs" -- also has parsimony on its side. The human mind, like the rest of the body and its functions, was designed by natural selection. The more important the adaptive problem, the more intensely selection will have specialized and improved the performance of information processing mechanisms for solving it. Domain general information processing mechanisms simply cannot insure adaptive responses in evolutionarily important domains of human activity -- domains like social exchange. Reasoning in such domains should be governed by

"Darwinian algorithms": mental algorithms specialized for solving the adaptive problems that define these domains.

With this thesis, I hope to resurrect the arguments against equipotential psychological mechanisms. From the standpoint of evolutionary theory, nothing could be more unparsimonious than the view that the human mind is a general purpose information processor. Yet the application of the Chomskian view has been limited because cognitive psychologists have lacked a systematic heuristic for judging which domains, other than language, were likely to command functionally distinct mental organs. Because it is a theory of function, natural selection theory provides just such a heuristic. Evolutionary principles allow one to pinpoint domains for which natural selection can be expected to have shaped how humans reason. Moreover, they suggest computational theories (sensu Marr) of what their design features are likely to be.

The theory of social exchange developed in this thesis was informed, at every stage, by evolutionary principles. In the study of human reasoning, the search for content-independent inference procedures had generated a confusion of apparently contradictory results; the hypothesis that humans have domain specific Darwinian algorithms for reasoning about social exchange resolves much of this confusion. The heuristic application of evolutionary theory can revolutionize cognitive psychology, allowing it to address issues closer to the heart of what we think of as human nature. This thesis is offered as a small illustration of its potential.

Chapter 1

Logic and the Study of Human Reasoning

1.1 Why were psychologists interested in deductive logic?

> The world, in short, was providing not sensation but fodder for our hypotheses.
>
> -- Jerome Bruner*

For many, the degree to which human learning mechanisms can be counted on to produce valid knowledge is the measure of man's rationality. But what characteristics must a learning process have to ensure that the knowledge acquired is valid? Because the generalizations of science afford the closest approach to what we intuitively think of as valid knowledge, psychologists have watched the philosophy of science closely to learn which characteristics of the scientific learning process are epistemologically criterial. If everyday learning can be shown to share these criterial characteristics with its more refined sister, then human rationality is spared.

Ever since Hume, induction has carried a heavy load in psychology while taking a sound epistemological beating. In psychology, it has been the learning theory of choice since the British Empiricists argued that the experience of spatially and temporally contiguous events is what allows us to jump from the particular to the general, from sensations to objects, from objects to concepts. Pavlovian reflexology, Watsonian and Skinnerian behaviorism, even the sensory-motor parts of Piagetian structuralism have been mere essays on the inductive psychology

---

* Bruner, 1984, p. 95, emphasis his.

of the British Empiricists. Yet when Hume, a proponent of inductive inference as a psychological learning theory, donned his philosopher's hat, he showed that induction could never justify a universal statement. Thus, Hume showed that the process by which people were presumed to learn about the world could not ensure that the generalizations it produced would be valid.

Only recently, with the publication in 1959 of Karl Popper's The Logic of Scientific Discovery, has the philosophical beating of psychology's favorite learning theory subsided. Popper argued that although a universal statement of science can never be proved true, it deductively implies particular assertions about the world -- hypotheses -- and particular assertions can be proved false. No number of observed white swans can prove that "All swans are white" is true, but just one black swan can prove it false. Generalizations cannot be confirmed, but they can be falsified, so inductions tested via deductions are on firmer epistemological ground than knowledge produced through induction alone.

This view had consequences for psychologists interested in everyday learning. If one assumes that the evolutionary purpose of human learning is to produce valid generalizations about the world, then surely everyday learning must be some form of Popperian hypothesis testing. On this view, induction still plays an important role in the creation of knowledge -- it is a source of testable hypotheses. But the burden of validating knowledge now falls on deductive logic. Inductive processes might suggest "If P then Q" -- and "If R then Q", "If S then Q", and so on -- but it is our use of deductive logic that causes us to reject the hypothesis if Q turns out to be false when P (or R

or S) is true. The repeated application of deductive logic in testing the many inductively produced hypotheses explaining Q lets us hone down the possibilities and zero in on the truth. On this view, human learning processes will produce valid knowledge to the extent that they use deductive logic to falsify hypotheses.

This shifts theoretical priorities in psychology from the study of inductive processes to the study of deductive logic. Either everyday human learning mechanisms usually produce invalid knowledge, or they include algorithms that frequently and spontaneously apply deductive logic in testing hypotheses.

A legion of cognitive psychologists, from Piaget (e.g., Inhelder & Piaget, 1958, p. 254-255) to Bruner (Bruner, Goodnow & Austin, 1956) to Wason & Johnson-Laird (1972) to Fodor (1975) have adopted this Popperian view of hypothesis testing as their model of human learning. To paraphrase Fodor, who was speaking for the field, hypothesis testing is the only theory we've got (Fodor, 1975, Ch. 1).

Deductive logic has another property that was tempting to cognitive psychologists: it is content-independent. The rules of inference of the propositional calculus* generate only true conclusions from true premises, regardless of what the propositional content of those premises is. The propositional calculus is the perfect inference engine for a domain general information processing system: no matter what hypotheses inductive processes feed it, it will output only valid conclusions. The idea that the human mind has algorithms that

* the philospher's name for formal propositional logic (Quine, 1950).

instantiate the rules of inference of the propositional calculus fit well with cognitive psychology's meta-theory.

Consequently, many psychologists have spent a great deal of time and effort in search of a "deductive component" (Wason & Johnson-Laird, 1972), or, in more current parlance, a logic "module." The theoretical burden they have placed on this proposed mental algorithm is staggering: It is supposed to be necessary for building virtually all the vast and complex knowledge structures that power human thought and behavior, from the most ubiquitous of social interactions to the most esoteric feat of modern technology.

1.2 What would a logic module be like?

The "doctrine of mental logic" (Johnson-Laird, 1982) is the view that the human mind includes innate algorithms instantiating the rules of inference of the propositional calculus -- a logic module. What properties can a logic module be expected to have?

Chomsky (1975), Marr & Nishihara (1978), and Fodor (1983) have taken the biological view (best summarized by Williams, 1966) that if a function is evolutionarily important, natural selection will produce a species-wide psychological mechanism with certain properties. Namely:

1. It will be specially designed to solve the evolutionary problem quickly, reliably, and efficiently. Consequently, it will instantiate mental architecture and rules of inference that will define the evolutionarily salient dimensions of the problem, and guide the organism toward an adaptively appropriate solution.

2. It will be domain specific. Only by limiting its scope of application can it be specially designed to solve the problem quickly and efficiently. I would add that it must have design features that make it sensitive to cues that indicate when the organism has encountered the domain for which the mechanism was designed. An algorithm that allows you to decide between fight or flight in the presence of a predator is useless unless it has features that let you determine what counts as a predator and when you are in the presence of one.

3. It will develop without explicit teaching or training. Exposure to the domain may be necessary to activate the mechanism or to allow it to fill in parameter values. But the rules that organize and process the stimuli are innately specified.

4. The inferences will be made automatically, without the application of "conscious effort" or deliberation. This is a consequence of its having to be fast and reliable (to remain reliable, the rules must be protected from the effects of deliberation -- they cannot be "isotropic" (Fodor 1983)).

Following this view, a logic module necessary for generating vast knowledge structures ought to have several properties:

Criterion A. It should instantiate procedures that reliably lead to valid deductions. Otherwise, it would not let you reject invalid hypotheses, and that is its proposed function.

Criterion B. It should be able to "recognize" hypotheses (in or out of consciousness), and upon recognizing them, correctly

process them. This is because its domain is the universe of possible hypotheses.

Criterion C. It should process hypotheses quickly, automatically, and without "conscious attention." There are an infinite number of ways of carving the world into properties, and therefore an infinite number of relations between properties to serve as hypotheses; on average, an enormous number of hypotheses will have to be tested before the correct one is hit upon and the simplest generalization made (this point may be fatal to the entire learning-as-hypothesis-testing view). Therefore, processing must be quick and automatic.

Criterion D. It should develop without any special teaching. Adults rarely sit down and teach children the canons of formal logic, yet children learn things constantly. This means one of two things: either deduction is not necessary for most learning, or the logic module (or a "logic module acquisition device"!) is innate. If the logic module is necessary for learning, then it itself cannot be learned (Johnson-Laird, 1982). Hence, supporters of a hypothesis-testing view of learning are committed to an innateness position (whether they realize it or not).

Criterion E. With respect to the propositional content of the hypotheses it processes (what P and Q stand for in "If P then Q"), it should be content-independent. Because this mechanism is supposed to account for learning in all domains, the domain from which the propositional content of the hypothesis is taken should have no effect on how quickly the deduction is

made or how likely it is to be valid.

The "doctrine of mental logic" was inspired primarily by criteria A and E. The propositional calculus is a system of rules for the derivation of valid inferences from propositions linked by logical connectives like and, not, and or. A logic module instantiating the propositional calculus would therefore satisfy criterion A, that the module instantiate procedures that reliably lead to valid deductions. In addition, conclusions derived via the propositional calculus are valid regardless of the specific content of the propositions involved. Its rules depend only on the truth values assigned to the propositions (whether the individual propositions are considered true or false) and on their position with respect to the logical connectives. For example, the "truth tables" associated with a conditional statement (If P then Q) and a biconditional statement (P if and only if Q) are:

Conditional

P	Q	If P then Q
T	T	T
T	F	F
F	T	T
F	F	T

Biconditional

P	Q	P iff Q
T	T	T
F	F	T
T	F	F
F	T	F

Thus, if P and Q are both considered true, then "If P then Q" is also considered true. Therefore, if P stands for "the sea is blue" and Q stands for "quantum physics is correct" -- and both these statements are considered true -- then the statement "If the sea is blue then quantum physics is correct" is also considered true. This property of the propositional calculus

satisfies criterion E, that the logic module's rules of inference be content-independent.

1.3 Do humans have a logic module?

Above, I argued that if learning occurs through Popperian hypothesis testing, then humans must have a logic module -- a psychological mechanism with algorithms that allow people who have had no special training in logic to recognize hypotheses and deduce only their valid implications, quickly, reliably, and automatically. Research on deductive reasoning indicates that humans have no such ability (for reviews, see Wason & Johnson-Laird, 1972; Johnson-Laird, 1982). Because there is so little dissent on this point among psychologists who study logical reasoning, I will cite only a few illustrative examples, drawn from the literature on reasoning about conditional statements.

In reasoning about conditional statements, one can make two correct inferences and two fallacious inferences (to convince yourself, inspect the truth table in section 1.2):

Correct inferences		Fallacious inferences	
Modus ponens	Modus tollens	*Affirming the Consequent	*Denying the Antecedent
If P then Q	If P then Q	If P then Q	If P the Q
P	not-Q	Q	not-P
Therefore Q	Therefore not-P	Therefore P	Therefore not-Q

* These inferences are fallacious because a conditional does not claim that P is the only possible antecedent of Q. Consider a concrete, causal statement: "If it rains then the grass is wet." If it has not rained, the grass may or may not be wet -- perhaps I have been watering the lawn with my sprinkler. To conclude "it rained" (or "it did not rain") from the rule premise and the "grass is wet" premise is to "affirm the consequent". To conclude "the grass is dry" (or "the grass is wet") from the rule premise and the "it did not rain" premise is to "deny the antecedent". No valid inference can be drawn from these sets of premises.

Minimally, a logic module capable of evaluating conditional hypotheses should instantiate procedures that quickly and reliably accomplish modus ponens and modus tollens. Furthermore, it should be immune to the two fallacious inferences. The algorithms involved are simple and well-defined -- a microcomputer can easily be programmed to run them.

Moreover, the only fair way to test for a logic module is to use statements that express unfamiliar relations, such as "If an object is a triangle, then it is red", or "If there is an A on one side of the card, then there is a 3 on the other side". The logic module is supposed to be necessary for learning, that is, for the construction of <u>new</u> knowledge. If this is its purpose, then it should be good at handling unfamiliar relations. Furthermore, the use of relations drawn from unfamiliar domains provides a cleaner experimental design. It prevents subjects from simply "looking up facts" to answer the question: one need not engage in any reasoning process to decide that "All swans are orange" is false. Conditionals relating letters to numbers are good candidates because letters and numbers are familiar enough but relations between them are not. For lack of a better term, I will follow the literature and call such conditionals "abstract."

Prediction A: Valid deductions are made frequently and reliably.

Item. Shapiro (reported in Wason & Johnson-Laird, 1972, pp. 43-44) asked 20 subjects to evaluate the validity of abstract versions of the four inferences listed above. If humans have a logic module, her subjects should make few if any errors: they should judge the first two inferences valid and the last two

inferences invalid. Errors should be randomly distributed among the four inferences. This task is very simple -- it does not even require subjects to generate conclusions themselves. All they have to do is correctly recognize inferences that have already been made as valid or invalid.

The error rate was reasonably low for modus ponens (5%), but the error rate was 52.5% for modus tollens, 20% for affirming the consequent, and 25% for denying the antecedent. Half the time subjects were judging a valid inference invalid, and a quarter of the time they were judging invalid inferences valid. Errors were not randomly distributed among the four conditions. The distribution of errors indicates that subjects find it particularly difficult to recognize the validity of modus tollens.

Item. In an experiment by Gibbs (reported in Wason & Johnson-Laird, 1972, p. 57-59) subjects had to generate deductions. On average, 44% of the problems requiring the use of modus ponens were done incorrectly, and 80% of those requiring modus tollens were done incorrectly. In both cases, incorrect inferences corresponded to committing the fallacy of affirming the consequent. Modus ponens was correctly used 2.8 times as often as modus tollens was.

Item. Mazzocco (reported in Legrenzi, 1970) found that subjects erroneously assume that "If P then Q" is equivalent to "If Q then P" when this makes a problem easier to "solve". Pollard & Evans (1980) found that subjects frequently view logically distinct conditionals as implying one another.

Item. Pollard & Evans (1981) found that subjects have a

pronounced tendency to judge an inference valid when they agree with the conclusion, and invalid when they do not agree with the conclusion -- regardless of its true validity.

The claim that humans have a quick and reliable deductive component seems to fall before it takes its first step. Experimental results do not even support criterion A, that people be able to reliably make valid deductions. Although people have some measure of success in recognizing the validity of modus ponens, they are not good at using it to generate deductions. They are quite susceptible to making fallacious inferences, and they seem to lack a procedure corresponding to modus tollens almost entirely. The literature on logical reasoning is quite consistent on this point. According to Johnson-Laird (1982), "the doctrine of logical infallibility is either falsified by the results of some experiments on syllogistic reasoning or else empirically vacuous."

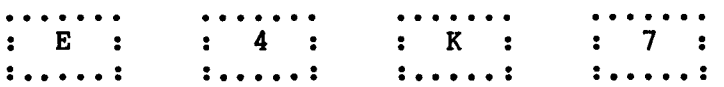
To save this perspective, one might argue that the logic module has a simpler design and a more specific function. Perhaps it does not have procedures for deriving deductive implications at all: perhaps it can only look for falsification. Nothing could be simpler to program. Consider any hypothesis of the form, "If P then Q." The truth table for the conditional shows that there is only one circumstance that can falsify this hypothesis: the co-occurrence of P and not-Q. A logic module capable only of falsification would scan all instances of P and all instances of not-Q. It would reject the hypothesis if any P was paired with a not-Q or if any not-Q was paired with a P.

The Wason selection task tests this prediction. Peter Wason was interested in Popper's view that the structure of science was hypothetico-deductive. The selection task allows one to see whether people really are falsificationists in testing hypotheses. In the selection task, a subject asked to test a hypothesis of the form "If P then Q" with respect to a universe of four cards representing possible pairings of P and not-P with Q and not-Q. Here is the original selection task (Wason, 1966):

Consider the following sentence:

"If a card has a vowel on one side then it has an even number on the other side."

It refers to these four cards:



Each card has a letter on one side and a number on the other side. Name those cards, and only those cards, which need to be turned over in order to determine decisively whether the sentence is true or false.

The cards were real cards, and an experimenter administered the task in person to one subject at a time.

The Wason selection task has a general solution. Turn over all cards displaying P (to see if they have a not-Q on the other side) and turn over all cards displaying not-Q (to see if they have a P on the other side). There is no point in turning over Q or not-P, because any value on the other side of these cards would be consistent with the hypothesis.

This provides a direct test of the modified view of the logic module's function. If the logic module is specialized for testing hypotheses through deductive falsification, then subjects

should immediately realize that they must turn over the E card (P) and the 7 card (not-Q).

They do not. On average, only 4 to 10 percent of all subjects choose P and not-Q when confronted with an abstract hypothesis (Wason, 1983). The majority pick only P, or P and Q, as if they are trying to confirm the existence of a relation, rather than falsify a proposed relation. This result has been replicated many times, under a wide variety of conditions: with different abstract propositions standing in for P and Q, with variations in the linguistic format of the hypothesis, with variations in how the information is represented on the cards, with variations in the instructions (e.g., Wason, 1968; Wason, 1969a & b; Wason & Johnson-Laird, 1970; Wason & Shapiro, 1971; Goodwin & Wason, 1972; Wason & Golding, 1974).

Furthermore, it is very difficult to teach people the solution. A wide variety of "therapies" have been tried; they have resulted in little or no facilitation in falsification rates. For example:

1. For each of 24 sample cards, subjects were asked whether or not each card was consistent with the rule; they were given feedback about their answers. They were then asked to solve a selection task using the same rule (Wason & Shapiro, 1971).
2. For each of 24 sample cards, subjects were asked to imagine a value on the other side of the card that would falsify or verify the rule; they were given feedback about their answers. They were then asked to solve a selection task using the same rule (Wason & Shapiro, 1971).

3. Subjects were allowed to turn over the cards they had selected, asked whether each verified or falsified the rule, and corrected if wrong; then they were retested (Hughes, 1966).
4. A duplicate set of fully revealed cards were present for subjects to inspect (Goodwin & Wason, 1972).

Even professional logicians have been known to get the problem wrong! (Wason & Johnson-Laird, 1972, p. 179)

Subjects' performance on the Wason selection task is the most damning evidence against the learning-as-hypothesis-testing view. Confronted with a novel hypothesis, subjects do not try to falsify it. Yet this is a paradigmatic case in which they should use deductive falsification. This result falsifies the modified view of the logic module as specialized for spotting falsifying evidence. In addition, because modus ponens and modus tollens can be used to solve the selection task, this result, like the previously cited evidence, falsifies the original view of the logic module as instantiating deductive procedures to be used in falsifying hypotheses.

Beating a dead horse.

A logic module necessary for learning should meet four other criteria (B-E), but these are predicated on it fulfilling prediction A -- that people frequently and reliably make valid deductions. Prediction A has been shown to be false, so technically, the other predictions fall with it. Just to be thorough, however, I would like to briefly discuss each separately.

Prediction B: The logic module can "recognize" hypotheses, and upon recognizing them, process them.

Hypotheses about the world do not come in just one linguistic format. A logic module should be able to recognize and operate on the logical "deep structure" of a hypothesis, producing valid deductions regardless of its linguistic format. The amount of time the conversion to deep structure takes may differ with linguistic format, not the validity of the deductions made.

This is not the case. A number of studies show that (1) different linguistic formats of the same hypothesis differ in how likely they are to elicit a valid deduction, and (2) a linguistic format that facilitates deduction for one problem may impede deduction in another (e.g., Van Duyne, 1974; Roberge, 1978, 1982; Bracewell & Hidi, 1974). Subjects in these studies had no time constraints, so differences in performance can be accounted for only by differences in linguistic format.

Prediction C: Valid deductions are made quickly, automatically, and without conscious attention.

In the experiments cited under prediction A, subjects were permitted to devote all the time and conscious attention to the problem that they wanted, yet they still did not make valid deductions. Clearly they do not make valid deductions quickly, automatically, and without conscious attention.

Prediction D: The logic module develops without any special teaching.

Again, the evidence cited for prediction A shows that people do not reliably make valid deductions without special training. Indeed, it is not clear that they reliably make valid deductions

even with special training. As the therapy experiments showed, performance on the Wason selection task proved relatively impervious to special training techniques, and even professional logicians find it difficult.

Prediction E: The logic module is content-independent.

Wason's first selection tasks used hypotheses that expressed abstract relations, usually involving letters and numbers. Performance was uniformly poor. However, a number of experiments in the early 1970's reversed this result (Wason & Shapiro, 1971; Johnson-Laird, Legrenzi & Legrenzi, 1972; Bracewell & Hidi, 1974; Gilhooly & Falconer, 1974). These experiments suggested that if the content of the rule being tested expresses a "familiar," "realistic," or "thematic" relation, subjects do reason logically on the selection task. This enhancement of logical performance with familiar materials is known as the "content effect" or the "thematic materials" effect on the Wason selection task.

Initially, researchers thought that the familiarity or realism of thematic content somehow facilitates the use of deductive logic (Wason & Shapiro, 1971; Johnson-Laird, Legrenzi & Legrenzi, 1972). The problem with this explanation is that the phenomenon is quite difficult to replicate. Some familiar content seems to facilitate the use of deductive logic; other familiar content does not (e.g., Van Duyne, 1976; Manktelow & Evans, 1979; Griggs & Cox, 1982; Cox & Griggs, 1982; Reich & Ruth, 1982; Yachanin & Tweney, 1982; Griggs & Cox, 1983). In addition, the same familiar content seems to facilitate logic at some testing locations, but not at others (e.g., Golding, 1981;

Griggs & Cox, 1982; Yachanin & Tweney, 1982). This should not happen if familiar content simply activates a logic module.

Whatever the explanation, the cognitive processes that govern reasoning about logical conditionals in the Wason selection task are clearly not content-independent.

The hypothesis that humans have the sort of logic module necessary for Popperian-style everyday learning faltered before taking its first step. Not one of the five defining criteria of a logic module is fulfilled by the results of experiments on logical reasoning.

This raises some serious questions: If people are not using deductive rules to reason about conditional statements, then what rules are they using? And if people are not learning via Popperian hypothesis testing, then how are they learning?

The discovery of the content effect on the Wason selection task raises the possibility that reasoning about logical conditionals is governed by content-dependent cognitive processes. Indeed, after years spent researching this effect, Wason and Johnson-Laird commented that the conditional "is not a creature of constant hue, but chameleon-like, takes on the colour of its surroundings: <u>its meaning is determined to some extent by the very propositions it connects</u>" (1972, p.92, italics theirs). They say that the principles governing the "cohesion of discourse" probably hold the key to its many meanings, and that "the nature of these principles is little understood -- they probably involve more than purely linguistic factors." The investigation of Darwinian algorithms presented in this thesis is a preliminary enquiry into what "more" they involve.

CHAPTER 2

A review of the literature on the "elusive" content effect on the Wason selection task

When content effects are found, the hypothesis that they were produced by content-dependent cognitive processes should be entertained. The extensive literature on the Wason selection task is replete with reports of content effects. If content-dependent inference procedures exist, this literature is a promising place to look for them.

Attempts to predict and explain content effects on the Wason selection task in terms of "differences in subjects' experience" with the different content domains tested have created a hornet's nest of apparently contradictory results. The unpredictability and unreplicability of the content effect on the Wason selection task is so pronounced that Peter Wason has called it a "crisis" (personal communication) and Griggs and Cox (1982) have dubbed the effect "elusive."

Because the explanation of this elusivity is the subject of my thesis, this chapter will explore these results in some detail, one content area at a time. If domain specific reasoning processes are involved, then the data should resolve into patterns when it is categorized by content domain per se, but not when it is categorized by factors correlated with content, like "familiarity" or "realism." Five major content areas have been explored in the literature: transportation, food, school, and "social contracts."

The discussion of published explanations attempting to

account for these results is deferred to the next chapter, so they can be discussed in light of all the reported data.

2.1 The Transportation Problem

The "Transportation Problem", developed by Wason & Shapiro (1971), has been used in more experiments testing for an effect of thematic content in the Wason selection task than any other thematic rule. It is a conditional rule linking a place to a means of transportation, for example, "If I go to Boston, then I take the subway." Researchers always use places and means of transportation that are local for and familiar to their subject population. There are nine experiments comparing performance on the transportation problem to performance on an abstract problem. Two found substantial content effects, two found weak content effects, and five found no content effects at all.

Wason & Shapiro, 1971

The first demonstration of a content effect was by Wason & Shapiro (1971). They gave 16 subjects a selection task using the rule "Every time I go to Manchester I travel by car" (thematic group), and 16 subjects a selection task using the rule "Every card which has a vowel on one side has an even number on the other side" (abstract group).* Destinations and means of transportation were rotated in the thematic group to avoid the possibility of an effect due to preconceptions about the relation between particular destinations and means of transport. Sixty-two

* Only Wason & Shapiro's abstract problem used the vowel-even number rule. Abstract problems in the other studies linked specific letters and numbers, e.g., "If there is an 'A' on one side of a card, then there is a '3' on the other side."

percent of the thematic group gave the logically correct, falsifying answer, 'P & not-Q', whereas only 12% of the abstract group gave this answer (phi = .52).

Bracewell & Hidi, 1974

In 1974, Bracewell & Hidi and Gilhooly & Falconer tried to replicate Wason & Shapiro, 1971. Their experiments were designed to tease apart the relative contribution to success on the selection task of concrete terms versus concrete relations. Here I will only discuss the conditions that are directly comparable to Wason & Shapiro's thematic and abstract groups, because establishing the existence of a content effect is theoretically prior to asking what causes it.*

Noting that the most common selection task error is to incorrectly select the card corresponding to the Q term, Bracewell & Hidi wondered if subjects "spend more time analysing the first set of terms to the detriment of the second set." They tested this by framing their thematic and abstract problems in two different linguistic formats: "Every time P, Q" and "Q every time P" (e.g., "Every time I go to Ottawa I travel by car" and "I travel by car every time I go to Ottawa.") The logical structure of these two problems is identical, however the Q term comes first in the "Q every time P" format. Their results are pictured in Table 2.1.

The "Every time..." linguistic format (also used by Wason & Shapiro) successfully replicated Wason & Shapiro's findings: 9 out of 12 subjects (75%) answered 'P & not-Q' in response to the thematic rule versus 1 out of 12 (8%) for the abstract rule

* The relative contribution question will be considered in Chapter 3.

Table 2.1 Results of Bracewell & Hidi, 1974

	thematic	abstract	Totals
Every time P, Q:	9	1	10
Q every time P:	2	1	3
Totals:	11	2	

Number of subjects choosing 'P & not-Q'; n=12 per cell.

(phi = .68). However, there was no thematic content effect for the "Q every time P" phrasing; 2 out of 12 subjects in the thematic group falsified (17%), compared to 1 out of 12 (8%) in the abstract group. There is no reason to believe that this second phrasing is an unnatural one*; in fact, this is the phrasing which Bracewell & Hidi had hoped would enhance logical performance by focusing attention on the Q term.

Thus, a simple change in linguistic format completely erased the content effect.

Interpretation of Bracewell & Hidi's data is further complicated by the fact that they explicitly told their subjects that the conditional is not "reversible." This instruction is unprecedented in selection task research; so are its apparent effects. The most common response to abstract problems is usually 'P & Q'. Yet only 1 out of the 24 subjects in Bracewell & Hidi's two abstract conditions gave this response, and no one gave it in either of the thematic conditions.

This instruction is so serious a confound that some

* I would guess that pragmatic factors determine which phrase would come first in ordinary conversation -- whether the speaker wished to indicate that the topic of the sentence is going to Manchester ("Every time I go to Manchester...") or traveling by car ("I travel by car...").

researchers are hesitant to count Bracewell & Hidi's "Every time" condition as a replication of Wason & Shapiro (Manktelow, 1979; Griggs & Cox, 1982; Griggs, 1983). I believe it may have introduced demand characteristics of the following kind.

When stumped by a problem, people sometimes ask "what's the trick?" -- it is a request for insight into the problem. However, I have never heard anyone stumped by a problem ask, "what are the tricks?" In other words, people usually assume that a thought problem has one "trick", not two. But solving the selection task involves two "tricks": according to Wason & Johnson-Laird (1972), subjects have not achieved "complete insight" unless they realize 1) that the Q card is irrelevant, and 2) that the not-Q card is relevant. When I was conducting pilot experiments, subjects who had finished the task frequently asked me if the "trick" was realizing that one should omit Q. And, in fact, the second most common response on the selection task is 'P' alone, omitting the Q card.

If you believe you have found the "trick", why look for a second one? Telling subjects that the conditional is "not reversible" may be giving away half the game. When the meaning of "not reversible" is clear, this instruction is equivalent to telling them to omit the Q card. The task of finding "the trick" remains.

This could explain why subjects found the "not-Q trick" in the "Every time P, Q" format, but not in the "Q every time P" format. When the logical operator that defines the conditional is at the beginning of the sentence, as in "Every time I go to Ottawa I travel by car", the meaning of "reverse" is

straightforward. Subjects have virtually been told to omit Q, so they continue to search for the problem's trick -- choosing not-Q -- and may eventually find it. But the meaning of "reverse" is far more ambiguous when the logical operator is in the center of the sentence. What is the "reverse" of "I travel by car <u>every time</u> I go to Ottawa"? Is it "Every time I go to Ottawa I travel by car" or is it "I go to Ottawa every time I travel by car"? Figuring this out may have been challenging enough to count as "the trick" for subjects solving problems in the 'Q every time P' format. Telling them the conditional was "not reversible" was enough of a clue to allow most of them to finally figure out that they were supposed to omit Q, but having realized that, they stopped their search -- they thought they had found the "trick."

If we leave aside this methodological objection, Bracewell & Hidi's experiments can be thought of as two separate attempts to replicate Wason & Shapiro: one a success, the other a failure.

Gilhooly & Falconer, 1974

The design of Gilhooly & Falconer was similar to Bracewell & Hidi, except they used only the "Every time P,Q" linguistic format, and many more subjects (n=50 per group).

Gilhooly & Falconer's thematic group did significantly better than their abstract group: 22% v. 6% chose 'P & not-Q'.* However, the success rate for the thematic condition was quite low: 22%, as compared to 62% for Wason & Shapiro and 75% for

* Gilhooly & Falconer were puzzled that the error responses which Johnson-Laird & Wason (1970) classify as "partial insight" ('P, Q, and not-Q') did not have the same distribution as the "complete insight" ('P & not-Q') responses. If one grants their assumption that these two scores express progressively greater degrees of insight into the logical structure of the problem, and therefore lumps them together, the content effect disappears (26% v. 18%)

Bracewell & Hidi's best group. The effect size, phi = .23, for Gilhooly & Falconer is closer to the "effect" size of .13 for Bracewell & Hidi's no effect condition than it is to the phi of

for Wason & Shapiro. It is not unheard of for 22% of a subject population to get the abstract problem correct; in a number of my experiments, more than 22% of subjects falsified on the abstract problem (see Chapter 6). Furthermore, because Gilhooly & Falconer's sample size is three times larger than either Wason & Shapiro's or Bracewell & Hidi's, one might expect their figures to be somewhat less subject to Type 1 errors.

Thus, if one counts any facilitation with thematic content, however small, as a "content effect", then Gilhooly & Falconer counts as a replication of Wason & Shapiro. However, if by "content effect" one means that a majority of subjects give a logically correct response with a thematic rule, then Gilhooly & Falconer have failed to replicate Wason & Shapiro. As will be discussed in the next chapter, the theoretical claim being made determines which definition is appropriate.

Pollard, 1981

In a very close replication of Wason & Shapiro's initial study, Pollard (1981) found a mild content effect: 4 out of 12 subjects (33%) in the thematic condition gave the logically correct answer, whereas none of the 12 subjects in the abstract condition gave this answer (phi = .45).

It is worth noting that given the percentage difference between the two groups (33%), zero correct in the abstract condition is the only outcome that could yield a significant result. If the same percent difference is maintained, but just

one subject in the abstract condition had answered correctly (hence 1 out of 12 in the abstract group, compared to 5 out of 12 in the thematic group), the difference between the two conditions would be insignificant ($p < .07$, Fisher's Exact). Given Pollard's small sample sizes, such a precarious result should be interpreted with caution.

This, so far, has been the good (and lukewarm) news for the content effect with a transportation problem. Now, the bad news.

Manktelow & Evans, 1979

In 1979, Manktelow & Evans conducted an experiment (Experiment 5) that duplicated Wason & Shapiro (1971) in every respect except one: they used an "If-then" linguistic format instead of the "Every time" format used by Wason & Shapiro (1971), Bracewell & Hidi (1974), Gilhooly & Falconer (1974), and Pollard (1981). Performance for the thematic and abstract groups was identical.

Brown, Keats, Keats, & Seggie, 1980

Brown, Keats, Keats, & Seggie (1980) also tried to replicate Wason & Shapiro, 1971, using 24 Australian and 24 Malaysian university students. Like Wason & Shapiro, their transportation problem used an "Every time" linguistic format. Their abstract problem used shapes and letters: "Every card with a black triangle on one side has a Y on the other side." For Malaysian subjects, the selection task was translated into Bahasa Malaysia, their national language. For both problems, subjects were told that the variables were strictly binary (travel is only by car or airplane, trips are only to Singapore or Penang), thus reducing the array of possible combinations of values from an infinite set

to four. Unlike other transportation problem experiments, Brown et al. did not rotate destinations and means of transportation. Half the Australians and half the Malaysians were given the transportation problem; the remaining subjects were given the abstract problem. Brown et al. found no enhancement of logical performance with thematic content. None of the Malaysian subjects answered either problem correctly, one Australian answered the abstract problem correctly, and two answered the thematic problem correctly (the between-cultures factor was not significant).

Yachanin & Tweney, 1982

Yachanin & Tweney (1982) looked in vain for a thematic content effect in a variety of different content areas. Evans & Lynch (1973) argued that performance on selection tasks with abstract rules is guided by a "matching bias": a tendency to choose cards that match values mentioned in the rule, regardless of their logical status (i.e., regardless of whether the propositions in the rule are affirmative or negative). This can only be tested by systematically negating components of a rule. Thus, given the rule "If not-A then not-3", subjects would choose the "A" and "3" cards because they are directly mentioned in the rule. By coincidence, choosing the "3" card is "logically" correct because it represents a false consequent (not-Q), and choosing the "A" card is logically incorrect because it represents a false antecedent (not-P). The matching bias is considered a non-logical response bias because it is blind to the logical structure of the problem.

Yachanin & Tweney argued that if thematic content truly

facilitates logical reasoning, then it should "protect" subjects from matching bias. Hence, they used four forms of each "If-then" rule: 1) affirmative antecedent and consequent (AA), 2) negative antecedent and consequent (NN), 3) affirmative antecedent and negative consequent (AN), 4) negative antecedent and affirmative consequent (NA). Subjects were tested on two of each of these rule forms (a total of eight problems per subject). A subject's eight problems were either all thematic or all abstract (n=40 per group). Yachanin & Tweney found no difference in performance between their thematic group and their abstract group for any of the rule forms (transportation: 13%, abstract: 11%). They did find evidence for both matching bias and a verification strategy in both groups.

One could argue that this result is uninteresting because there is evidence (reviewed in Wason & Johnson-Laird, 1972) that negatives are difficult to understand, hence subjects simply became confused in this experiment. There are two problems with this criticism. The first is theoretical: Many explanations of why there should be a thematic content effect are based on the idea that, by virtue of their familiarity, imageability, coherence, etc., thematic rules make confusing statements easier to understand. Thus, one would still expect a relative enhancement for thematic rules with simple scope negative components when compared to abstract rules with the same structure of negation, even if performance on these thematic rules is not as high as performance on affirmative rules. The second problem with this criticism is empirical: Even when one considers only affirmative (AA) rules, there is no difference in

card selections between abstract and thematic groups. One would have to argue that merely being exposed to a rule with a negative component is sufficient to totally confuse subjects when they then encounter an AA rule. Manktelow & Evans (1979) tested this possibility and found no evidence for it (see section 2.2 below).

Griggs & Cox, 1982

In 1982, Griggs & Cox tried to replicate Wason & Shapiro's result (Experiment 1). They used 32 subjects and gave each two problems: a transportation problem and an abstract problem. Half got one first, half the other. Like Wason & Shapiro, they used the "Every time" phrasing. Unlike Wason & Shapiro, they found no difference in performance between the two problems.

Transportation Problem Summary

The transportation problem elicited a substantial (greater than 50% falsification rate) content effect in two experiments (Wason & Shapiro, 1971; Bracewell & Hidi, 1974), a weak content effect in two experiments (Gilhooly & Falconer, 1974; Pollard, 1981), and no content effect at all in five experiments (Bracewell & Hidi, 1974; Manktelow & Evans, 1979; Brown et al., 1980; Yachanin & Tweney, 1982; Griggs & Cox, 1982).

2.2 The Food Problem

The "Food Problem" was developed by Manktelow & Evans (1979). It is a conditional rule about meals, linking something a person eats with something that person drinks, for example, "If I eat salad then I drink water." Performance on the food problem has been compared to performance on an abstract (or "low thematic") problem in six experiments, four by Manktelow & Evans

(1979), one by Yachanin & Tweney (1982), and one by Reich & Ruth (1982). No one has found an enhancement in logical performance with the food problem.

Manktelow & Evans, 1979

Manktelow & Evans conducted four experiments using food problems (Manktelow & Evans, 1979, Experiments 1-4). The protocol for their Experiment 1 was similar to that described in section 2.1 for Yachanin & Tweney (1982). They systematically varied which logical component was affirmative or negative. Each subject was given an AA, AN, NA, and NN problem. Like Yachanin & Tweney, Manktelow & Evans reasoned that if thematic content facilitates the use of deductive logic, then subjects will be less likely to fall victim to the matching bias when given a food problem than when given an abstract problem. Every subject was given a test booklet with written instructions and four problems: 24 subjects were given four food problems, 24 were given four abstract problems. The 48 subjects were tested at the same time, as a group.

Performance on the food problems was as low as performance on the abstract problems, and both groups showed evidence of the matching bias. This result holds even if one considers only affirmative (AA) rules.

Puzzled by this result, Manktelow & Evans systematically varied task factors that could have interfered with logical performance. Experiment 2 was identical to Experiment 1, except subjects were tested individually, alone in cubicles, rather than in a group. The results were the same as for Experiment 1. Testing subjects individually or in a group appears to have no

effect on their performance.

Next, Manktelow & Evans wondered if presenting so many rules, and rules with some negated components, was simply imposing too great a "cognitive load" on their subjects -- confusing them. So in Experiment 3, each of 32 subjects answered only one, affirmative (AA) selection task (half were given a food problem, half an abstract problem). Subjects were run in small, unsupervised groups. Again, there was no difference in card selections between the thematic and abstract groups.

Last, Manktelow & Evans wondered if previous enhancements in performance with thematic problems could have been due to the presence and participation of the experimenter. In most of the earlier studies, the experimenter had read the instructions aloud, allowed subjects to inspect a deck of sample cards from which the four cards for the selection task were randomly drawn, and requested and recorded subject responses. Manktelow & Evans' Experiment 4 was identical to their Experiment 3, except subjects were run as described above, with the experimenter controlling the whole sequence of events. Again, there was no difference in performance between the thematic and abstract groups.*

One last point: Using Manktelow & Evans' data on the frequency with which individual cards were chosen for thematic and abstract groups,** one can consider the hypothesis that thematic content reduces confusion, even if it does not facilitate

---

* Manktelow & Evans' Experiment 5, using the transportation problem (described in 2.1), was also conducted this way. The results were the same.

** They report the frequency with which individual cards were chosen, regardless of the combination in which they were chosen.

logic by increasing the probability that not-Q is chosen. Manktelow & Evans had no hypotheses regarding the direction of differences for the P, not-P, and Q cards, so for these they used two-tailed Fisher's Exact tests. But suppose thematic content reduces confusion, and choosing not-P or Q, or failing to choose P, is evidence of confusion. Then one would use a one-tailed test with the prediction that not-P and Q are chosen less frequently for thematic problems and P is chosen more frequently. Their data do not support this hypothesis. Using one-tailed tests, there are no cases of differential choosing of P cards. In Experiments 1, 2, 4, & 5, there are no differences in the choice of not-P or Q cards between thematic and abstract groups. In Experiment 3 the thematic condition elicited fewer not-P choices ($p < .038$, predicted direction), but more Q choices ($p < .049$, opposite of predicted direction). Thus, Manktelow & Evans' data provide no support for the hypothesis that thematic content decreases confusion about the conditional's meaning.

Yachanin & Tweney, 1982

Yachanin & Tweney's (1982) study included a condition identical to their transportation problem condition (described in section 2.1), except that the thematic group was tested on food problems (the abstract group used for comparison was the same as that for the transportation problem). They found no significant difference in responses between the thematic and abstract groups (food: 14%, abstract: 11%). This is true even if one considers only the affirmative (AA) problems.

Reich & Ruth, 1982

For lack of a better place, I include Reich & Ruth (1982),

in the food problem section. Their experiment differs from the others in that they looked at performance on "low thematic" versus "high thematic" problems, without using an abstract problem for comparison. Their "low thematic" problems were food problems. Their "high thematic" problems were embedded in a story context, for example: "Molly is employed at a seaside cafe. She has been instructed by her boss to serve tea or coffee only at certain times of the day. Visitors notice that: When it is early Molly serves tea. Are they correct?" The object was to create a coherent, "unitary", easy to visualize scenario linking the terms of the conditional.

Like Yachanin & Tweney (1982), Reich & Ruth gave each subject one affirmative and three negated forms (AN, NA, NN) of each rule. Twenty-four subjects were given low thematic rules, 24 were given high thematic rules. High thematicity did not significantly improve logical performance, whether one considers all four rule forms (low thematic: 17%; high thematic: 22%) or only the affirmative form (low thematic: 4%; high thematic: 17%). Moreover, performance on the "low thematic" food problems was in the same low range of values typically found for abstract problems.

Food Problem Summary

None of the six experiments testing food problems elicited an enhancement in logical performance with respect to either abstract or "low thematic" problems. Furthermore, although some researchers (e.g., Pollard, 1981) have claimed that food problems are not as "thematic" as transportation problems, no one has yet proposed a criterion for judging "thematicity", nor has any one

produced a reasoned argument to support the claim that food problems are less thematic. Indeed, considering that people eat and drink at several meals every day, one might think that, if anything, food themes should be more familiar to subjects than transportation themes.

2.3 The School Problem

The school problem, developed by Van Duyne (1974), is a conditional relating a person's major field of study to his or her school, for example, "If a student studies philosophy, then he goes to Harvard." There are two experiments studying this problem: one found better performance for school problems than for abstract problems, the other did not.

Van Duyne, 1974

Van Duyne compared performance on abstract and school problems for four logically equivalent linguistic formats:

Universal: "Every student who studies physics is at Oxford."

Standard Conditional: "If a student studies philosophy then he is at Cambridge."

Disjunctive: "A student doesn't study French, or he is at London."

Conjunctive: "It isn't the case that a student studies psychology and isn't at Glasgow."

All four linguistic formats are logically equivalent to the conditional "If a student studies field A, then he goes to university B" (you can convince yourself of this by consulting the truth table for the conditional in Chapter 1: e.g., the disjunctive sentence is equivalent to "If a student studies French, then he is at London"). For all four linguistic forms,

the correct answer is to choose the major field mentioned in the problem and the university not mentioned in the problem.

Van Duyne made no attempt to rotate combinations of fields and schools to avoid effects due to preconceptions; the four sentences above are the four school problems he used. Each of his 24 subjects answered the four problems above and four abstract problems that had the same linguistic formats. Half the subjects answered the four abstract problems first, the other half answered the four school problems first.

For the abstract problems, there were no significant differences in percentage correct among the four linguistic formats. Performance on the disjunctive and conjuctive forms of the school problem was as low as performance on the abstract problems. However, there was a difference in performance between the school and abstract problems when they were phrased as universals and as standard conditionals. For the universal phrasing, 58% of subjects gave the logically correct answer on the school problem as opposed to 8% on the abstract problem. For the standard conditional phrasing, 50% answered correctly on the school problem, but only 12% on the abstract problem.

As in Bracewell & Hidi's (1974) experiment (section 2.1, transportation problem), the content effect disappeared in some linguistic formats. However, the absence of a content effect for Van Duyne's disjunctive and conjunctive formats is less damning than it is for Bracewell & Hidi's "Q every time P" format, which is a minor, and pragmatically common, variation on the universal format. Van Duyne's disjunctive and conjunctive formats contain a number of complicating confounds. For example, in English, "A

or B" can mean "A or B but not both" or "A or B or both". Also, his disjunctive's first component is a negative, and as mentioned previously, there is evidence that people have difficulty interpreting negatives (Wason & Johnson-Laird, 1972; of course, one could still argue that thematic content should lessen the interpretational difficulties). Furthermore, Van Duyne's conjunctive not only has two negatives, but it has two negatives of different scope -- the first is meant to encompass the whole following statement, whereas the second applies only to the school. Last, "A student doesn't study French or he is at London" is a rather bizarre way of saying "If a student studies French then he is at London," and "It isn't the case that a student studies psychology and isn't at Glasgow" is a strange way of saying "If a student studies psychology then he is at Glasgow." Pragmatically, a negative is usually used to contradict a presupposition that is the topic of conversation -- it is not used to introduce a topic (Clark & Clark, 1982, p. 99). For these reasons, I do not find the lack of a content effect for these two linguistic formats very interesting. Too many other factors could be swamping the effect.

The universal and standard conditional formats did elicit content effects. However, I would like to offer two caveats in interpreting this result.

1) Any of the subjects taking this test know that in real life the rule expressed by the school problem is false. Universities are not segregated by major field. It is simply false that all psychology students go to Harvard -- some go to

Yale, Tufts, U. Mass., etc., and, I presume, every student knows this. Compounding this problem, Griggs (1983) has pointed out that in the U.K., Cambridge is renowned for its excellence in physics (the rule pairs physics with Oxford), and Oxford is renowned for its excellence in philosophy (the rule pairs philosophy with Cambridge).

This creates interpretational problems because there is evidence indicating that if a subject has personal beliefs about the veracity of the relation expressed by a logical problem, that subject's performance on the logical problem is guided, in part, by a desire or tendency to confirm those personal beliefs. In other words, when subjects believe a statement to be true they try to verify it, and when they believe it to be false they try to falsify it (Janis & Frick, 1943; Wason & Johnson-Laird, 1972, Ch. 7; Van Duyne, 1976; Pollard & Evans, 1981; see Pollard, 1982 for review). Given the reputations of the schools used, a "belief bias" would lead to falsification. Thus, one doesn't know if the effect is due to belief bias, or due to the effect of thematic content as such. In testing for content effects, one wants to neutralize any effects of belief bias, not exacerbate them by using rules that are both false and contrary to the subject's personal prejudices.

2) This problem has some (but not all) of the earmarks of a social contract problem, for which there does seem to be reliable evidence for a content effect. Briefly, in social contract problems the conditional rule expresses a contract in which a person is eligible for a benefit if, and only if, she pays a price or meets a requirement (fuller descriptions follow in

section 2.4 and Chapter 5). Given such a rule in a Wason selection task, a subject looking for a "cheater" -- an individual who has absconded with the benefit without having paid the price or met the requirement -- would choose the same cards as a subject seeking logical falsification.

Van Duyne's school rule was embedded in a "look for cheaters" context. Subjects were told that the cards were taken from a register of students who are eligible for a grant (the benefit), and that certain rules are supposed to apply to eligible students (the requirements that must be met to get the benefit). Rather than being asked to turn over the cards necessary to see "whether the rule is true or not" (a common wording), subjects were asked to turn over the cards necessary to see "whether they violate the rule or not" (emphasis mine). "Violate" was used for both the abstract and school problems, so it cannot, in itself, explain the difference. But my point is that this choice of words in conjunction with a context defining eligibility for a benefit suggests that one is looking for a violator, that is, a cheater on a social contract. Thus, it was not clear to me whether I should include Van Duyne's school problem in this section or the next one, on social contract problems. According to the formulation that will be presented in Chapter 5, a full fledged social contract should have the benefit stated in the antecedent: in this problem, the benefit is stated in the context, and the rule states a conditional requirement that earning the benefit is contingent upon. Thus, it is a hybrid between a full-fledged social contract and a straightforward relational rule with thematic content. This

makes it difficult to know whether the effect was due to the use of thematic content in general, or whether it was specific to the use of a social contract context.

Please note that if this context exercised a major effect it would mitigate the criticism expressed in the first caveat, because the rule would not be interpreted as a statement about all students, merely about those interested in earning the grant.

Yachanin & Tweney, 1982

Yachanin & Tweney's (1982) study included a condition identical to their transportation and food conditions (described in sections 2.1 and 2.2), except that the thematic group was tested on school problems (the abstract group used for comparison was the same as that for the transportation and food problems). They do not say whether they rotated schools and major fields on different problems; they say only that their rules "were expected to be consistent with the experiences of the subject population."

Unlike Van Duyne (1974), Yachanin & Tweney found no significant difference in responses between the thematic school group and the abstract group (school: 12%, abstract: 11%). This is true even if one considers only the affirmative (AA) problems. Yachanin & Tweney do not report having embedded their rules in a story context, social contract or otherwise.

School Problem Summary

One study found a content effect with the school problem and one study did not. Van Duyne (1974) found a content effect with the school problem when it was presented in a universal or standard conditional format. He found no content effect when the school problem was presented in a disjunctive or conjunctive

format, but there are a number of confounds that could have swamped an effect for these linguistic formats. Because his school problems were embedded in a context that made his rule part of a social contract rather than a simple descriptive relation, and because his subjects were, in essence, asked to "look for cheaters", it is difficult to tell whether the facilitation he found is due to the use of thematic content in general, or due to the social contract content that it has. Yachanin & Tweney (1982), who did not embed their problem in a social contract context, found no enhancement in logical performance with the school problem.

2.4 Social Contract Problems

A social contract specifies what two or more individuals intend to exchange. In a social contract, whether an individual receives a benefit is contingent upon his paying a cost or meeting a requirement. Chapter 5 provides a detailed account of the structure of social contracts; my purpose here is to give the reader an intuitive grasp of this structure, so I can review the relevant literature.

A social contract rule relates perceived benefits to perceived costs, expressing an exchange in which an individual is expected to pay a cost to an individual or group in order to be eligible to receive a benefit from that individual or group. "Cheating" is the violation of a social contract rule; more specifically, cheating is the failure to pay a cost to which you have obligated yourself by accepting a benefit, and without which the other person would not have agreed to provide the benefit.

Cheating does not always correspond to logical falsification.

Consider the "Drinking Age Problem" (DAP; Griggs & Cox, 1982), pictured in Figure 2.1.

Figure 2.1

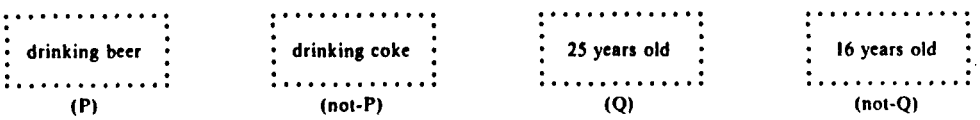
Drinking Age Problem (DAP; adapted from Griggs & Cox, 1982)

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

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

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

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



(P) (not-P) (Q) (not-Q)

For American subjects who perceive drinking beer as a rationed benefit that can only be had by waiting until they have met an age requirement (the cost), the DAP expresses a social contract of the form:

"If you take the benefit, then you pay the cost."

This same rule would not express a social contract to subjects who do not think of the right to drink alcohol as an age-rationed privilege. I am told that in the USSR, people of any age can buy and drink alcohol: it is a "free good" with respect to age. For Russian subjects the DAP rule would be merely descriptive, relating a predisposition for drinking beer to advancing age.* The transportation and food problems, as well as Yachanin & Tweney's school problem, were descriptive rules. Van Duyne's school problem was a prescriptive rule, but not a social contract

* Much as "If a person has a heart attack, then he must be over 20 years old" describes a relationship between advancing age and a predisposition to suffer heart attacks.

rule, because the benefit was not mentioned in the rule itself.

Figure 2.2 shows the structure of a Wason selection task that uses a social contract (SC) rule. Irrespective of logical category, a subject looking for potential cheaters should choose the "cost NOT paid" card (has he illicitly absconded with the benefit?) and the "benefit accepted" card (has he paid the required cost?).

Figure 2.2

Structure of Social Contract (SC) Problems

It is your job to enforce the following law:

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

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

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

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

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

Whether looking for potential cheaters on a social contract produces logically falsifying card choices on the Wason selection task depends on where the costs and benefits to the potential cheater are located in the "If-then" structure of the rule.

A "standard" social contract (STD-SC) is one where the benefit to the potential cheater is located in the antecedent clause and the cost/requirement is located in the consequent clause. Rule 1 of Figure 2.2 and the DAP are STD-SCs. For a STD-SC, the "benefit accepted" card corresponds to the logical

category "P", and the "cost NOT paid" card corresponds to the logical category "not-Q".

A "switched" social contract (SWC-SC) is one where the locations of cost and benefit are switched -- the cost is in the antecedent clause and the benefit is in the consequent clause. Rule 2 of Figure 2.2 is a SWC-SC. For a SWC-SC, the "benefit accepted" card corresponds to the logical category "Q" and the "cost NOT paid" card corresponds to the logical category "not-P".

Consequently, looking for cheaters on a STD-SC produces the logically falsifying, 'P & not-Q' response, whereas looking for cheaters on a SWC-SC produces a logically incorrect, 'not-P & Q' response.

In the search for content effects on the Wason selection task, 16 experiments have tested rules whose content expresses a standard social contract -- the format for which 'P & not-Q' is the choice of a subject who is looking for potential cheaters. A substantial content effect has been found in every one of these experiments.

2.4.1 The Post Office Problem

A post office problem is a conditional rule expressing a postal regulation, for example, "If a letter weighs two ounces, it must have 44 cents postage." Whether a particular post office problem is a social contract or not depends on the subject population or the problem's context. It is a social contract problem:

1. if its constituent propositions are recognizable as costs and rationed benefits to the subject population, or
2. if the story context surrounding the problem defines the

constituent propositions as costs and rationed benefits. However, if its constituent propositions are arbitrary with respect to the subject population's understanding of costs and rationed benefits, then the same rule is merely descriptive.

For Americans, the post office rule, "If a letter is sealed, then it must have 20 cents postage", is either descriptive or prescriptive,* but not a social contract, because sealing an envelope is not considered a rationed benefit in our culture -- sealing is just something one always does, a free good. However, this same problem is a social contract problem for older British subjects, because in Britain before 1968 one could pay lower rates if one left the letter unsealed. In other words, the privacy gained by sealing a letter was a benefit that had to be paid for in that culture.

Johnson-Laird, Legrenzi, & Legrenzi, 1972

Johnson-Laird, Legrenzi, and Legrenzi (1972) formulated the first and most famous post office/SC problem. Their thematic rule was:

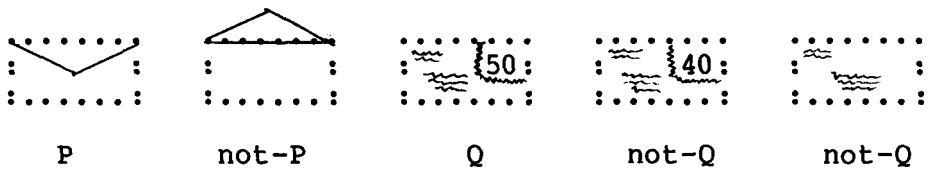
"If a letter is sealed, then it has a 50 lire stamp on it"

Their subjects were British. In Britain, within the memory of their subjects, sealing an envelope was a benefit rationed by ability to pay; unsealed envelopes could be sent at a lower rate (this is no longer the case). Furthermore, the subjects were instructed to imagine they were postal workers looking for letters that "violate the rule". Thus, the problem asked them to

* Depending on whether the person saying the rule is simply making an observation about sealed letters (descriptive) or telling you a seemingly arbitrary postal regulation (prescriptive).

look for cheaters on a social contract. They were then shown the following display of five real envelopes with real lire stamps:

Figure 2.3 Thematic problem "card" display, Johnson-Laird et al., 1972

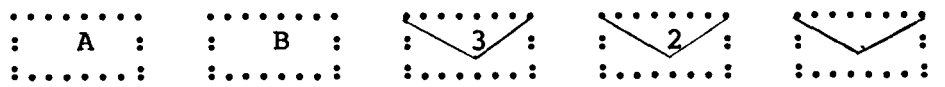


The envelopes are labeled with Ps and Qs for the reader's convenience; they were not so labeled in the actual experiment.

The logically correct answer is to choose 1) the sealed envelope (P), to see if the sender put too little postage, and 2) the 40 lire and no stamp envelopes (not-Q), which have too little postage to be eligible for sealing, to see if the sender illicitly sealed them. These are the same card choices a subject "looking for cheaters" would make.

Johnson-Laird et al. tested this rule in two different linguistic formats: "If P then Q" and "P only if Q" (e.g., "A letter is sealed only if it has a 50 lire stamp on it"). To reduce redundancy, one format used British stamps and the other use Italian stamps. Each of 24 subjects were given a total of four problems, one thematic and one abstract in each of the two linguistic formats. The order of problem presentation was randomized across subjects. The "cards" for the abstract problems were also real envelopes, with letters and numbers on the front and back. The rule reflected this by refering to the envelopes as "letters": "If a letter has an A on one side, then it has a 3 on the other side."

Figure 2.4 Abstract problem "card" display, Johnson-Laird et al., 1972



Ninety-two percent of the subjects gave the logically falsifying answer for at least one of the thematic problems, but only 29% gave this answer for at least one of the abstract problems. Seventy-one percent of the subjects got both thematic problems correct, but none of them got both abstract problems correct. There was no transfer from thematic problems to abstract problems, and when asked, only two of the 24 subjects realized that the logical structure of the thematic and abstract problems was similar. There were no effects due to linguistic format. Eighty-one percent of the 48 thematic problems (there were 2 per subject) were correctly solved, versus 15% of the abstract problems.

Golding, 1981

Golding (1981) conducted the Johnson-Laird, Legrenzi, & Legrenzi (1972) experiment using a population of British subjects ranging in age from 20 to 70 years. For subjects under 45 the post office problem was a prescriptive, non-SC problem, because the "lower rates for unsealed letters" rule was not in effect within their memory -- for them the constituent propositions were not recognizable as a cost and rationed benefit. Thus, there is no reason why these younger subjects would consider sealing an envelope a benefit that had to be paid for. However, this same

rule is a social contract problem for subjects over 45 who remembered the old regulation, because they would have recognized it as having a cost-benefit structure.

There was no difference in logical performance between the older and younger subjects on the abstract problem. However, 59% of the older subjects gave the logically correct, "look for cheaters" response to the post office problem, compared to only 9% of the younger subjects.

Golding herself framed the difference between the two groups in terms of familiarity rather than social contracts -- older subjects were more familiar with the rule than younger ones. In her view, older subjects perform better because they have direct experience with falsifying instances, and these instances are available in memory. In other words, older subjects do not reason about the rule, they simply recognize a pairing as falsifying if it occurs to them. There is nothing in her experiment to let one choose between her interpretation and a social contract interpretation.

Griggs & Cox, 1982

Griggs & Cox (1982, Experiment 2) conducted Johnson-Laird et al.'s post office problem using 24 American college students. For American students, the post office rule is not a social contract, because its constituent propositions are not recognizable as a cost and rationed benefit (sealing an envelope is a free good in the U.S.). Griggs & Cox's procedure differed from Johnson-Laird et al. in only two ways: 1) they used American and Mexican instead of British and Italian stamps, and 2) for the abstract condition they made clear the fact that letters could

appear only on the front of an envelope, whereas numbers could appear only on the back. Because Johnson-Laird et al. did not say this, the blank envelope in the thematic condition was clearly an instance of not-Q, whereas the blank envelope in the abstract condition could be categorized as either not-P or not-Q. Griggs & Cox argue that this is a serious confound because "two not-Q instances in the thematic arrays highlight a selection that is critical for the correct solution" (p.412).*

There was no significant difference in performance between the thematic and abstract problems (thematic: 1/24 correct, abstract: 4/24 correct). Furthermore, the confound discussed above cannot have accounted for Johnson-Laird et al.'s results, because performance on both thematic and abstract problems was low in this experiment. If Griggs & Cox's highlighting explanation had been true, their adding the not-Q highlight to the abstract problems would have raised abstract performance to the high level elicited by Johnson-Laird et al.'s thematic problems, not lowered thematic performance to the level of abstract.

Van Duyne, 1976

Van Duyne (1976) briefly reports the results of an as yet unpublished study ("Semantics and reasoning", in preparation) comparing the performance of four independent groups of subjects

* Johnson-Laird et al. could counter that the ambiguity is irrelevant because their rule refered to "one side of the envelope" versus "the other side"; hence, to be safe, one should assume the blank was a potential not-Q. However, this is an extra, rather subtle inference that one need not make in the thematic condition. Thus, Griggs & Cox's objection would still stand, as the thematic problem alone contains two clear, unambiguous not-Q instances, and the issue would have to be decided empirically, as it was.

on the following four rules (one problem per subject):

1. an abstract rule
2. a descriptive post office rule using real envelopes as "cards": "If there is L.B. Mill on one side of the envelope then there is PRINTED PAPER REDUCED RATE on the other side". This rule's constituent propositions are not recognizable costs and rationed benefits to British subjects.
3. an SC post office rule using cards to represent envelopes: "If there is PRINTED PAPER REDUCED RATE on one side of the envelope then it must be left open". This is a real postal rule in Britain, where the experiment was run. To be eligible for the lower rates for printed matter, the contents of an envelope cannot contain any personal correspondence. Thus, the lack of privacy needed for the post office to make spot checks is the price you pay to be eligible for the benefit of saving money on mailings of printed material.
4. the same SC post office rule as in (3), using envelopes as "cards".

He reports the following percentage correct for the four problems: (1) abstract -- 19.27%, (2) descriptive post office -- 48.96%, (3) SC post office (cards) -- 86.98%, (4) SC post office (envelopes) 97.92%. He says these four form a "highly significant trend", but does not say what statistical tests he performed, or even the number of subjects in the study.

So many details are omitted from his description of this study that its relevance is difficult to ascertain. However, it seems unlikely that his sample sizes could be so small that the average performance on the social contract problems of 92.45% and 86.98% would not be statistically different from 19.27% for the

abstract problem, or even from 48.96% for the descriptive post office problem. Assuming that Van Duyne used the usual scoring system ('P and not-Q' is a "hit", any other answer is a "miss"), the minimum number of subjects required to get the number "97.92%" for group (4) is 48 (47 out of 48 correct). It is doubtful, however, that his group sizes were equal. If they were, the minimum number of subjects per group required for all four percentages to be possible numbers is 192. As this would mean his study used a total of 768 subjects, I presume he had unequal group sizes. But if group sizes did vary around 48, the differences between all four groups would be strongly significant. If this were true, his descriptive post office condition would also have shown enhanced performance over the abstract condition. But there are too many ifs to judge.

2.4.2 The Drinking Age Problem

The "Drinking Age Problem", or DAP, was developed by Griggs & Cox (1982). It is a rule relating eligibility to drink alcoholic beverages to age, for example, "If a person is drinking beer then he must be over 20 years old" (see Figure 2.1). In our culture, drinking alcohol is a benefit that one is eligible for only after having met an age requirement, so the drinking age problem qualifies as an SC problem for American subjects.

Griggs & Cox, 1982

Each of 40 subjects, undergraduates at the University of Florida, were tested on a DAP and an abstract problem. Their DAP was: "If a person is drinking beer, then the person must be over 19 years of age." Nineteen was the legal drinking age in Florida

at the time of the experiment. Before the DAP, each subject was told: "Imagine you are a police officer on duty. It is your job to ensure that people conform to certain rules. The cards in front of you have information about four people sitting at a table..." and so on. For both thematic and abstract problems subjects were instructed to select the card(s) necessary to "determine whether or not the people are violating the rule." Half the subjects were given the DAP first, half were given the abstract problem first.

Seventy-two percent of the subjects gave the logically falsifying answer, 'P & not-Q', for the DAP, but no subject gave this answer for the abstract problem. There was no transfer from a correctly solved DAP to the abstract problem. Also, the percentage correct for each problem was the same, whether it came first or second.

Cox & Griggs, 1982

Cox & Griggs, 1982 (Experiment 1) replicated their finding of a content effect with the DAP (thematic: 60%, abstract: 4%). Furthermore, although they also replicated the lack of transfer from a correct DAP to an abstract problem, they did find transfer from the DAP to the "apparel-color problem" (ACP): "If a person is wearing blue then the person must be over 19." The ACP is a "semi"-social contract problem: its consequent is a culturally typical cost for American subjects -- we are familiar with age-rationed privileges -- but the antecedent is not a rationed benefit. Interestingly enough, the ACP elicited a sizable content effect only when it followed the DAP, a full-fledged social contract.

Griggs & Cox, 1983

To see if thematic content could "protect" subjects against matching bias, Yachanin & Tweney had presented subjects with thematic and abstract rules whose components had been systematically negated (AA, AN, NA, NN; see section 2.1). None of their thematic rules, including the AA rules, had facilitated logical performance. Griggs & Cox decided to reopen this issue, because by 1983 two thematic problems had been found that could reliably elicit falsifying responses from American subjects: the DAP and the Sears problem (see section 2.4.3 below). Each subject was given only one problem to solve. There were four thematic (DAP) groups and four abstract groups: one group for each rule form (AA, AN, NA, NN), 20 subjects per group.

Although Griggs & Cox (Experiment 2) systematically negated the components of the DAP, they preserved their structures as standard social contracts. The four thematic rules were:

AA: "If a person is drinking beer, then the person must be over 19"
AN: "If a person is drinking beer, then the person must not be under 19"
NA: "If a person is not abstaining, then the person must be over 19"
NN: "If a person is not abstaining, then the person must not be under 19"

Even though the components have been systematically negated, in each case the "If" clause refers to the benefit in question (drinking beer) and the "then" clause refers to the requirement that must be met to be entitled to that benefit (being over 19).

Control groups solved one of four similarly negated forms of Wason's (1966) original abstract problem, "If a card has a vowel on one side, then it has an odd number on the other side." With this rule, subjects can recode negated phrases as affirmative phrases (e.g., "not odd" to "even"), just as they can for the

DAP. This is not possible with ordinary letter-number rules.*

Negated conditionals are frequently difficult to understand (Wason & Johnson-Laird, 1972, see 2.1 on Yachanin & Tweney, 1982). Yet, in spite of the fact that three of the thematic SC rules had negated components, every SC rule elicited more 'P & not-Q' responses than its corresponding abstract rule (see Table 2.2).

Table 2.2 Griggs & Cox, 1983, Experiment 2 (DAP)

	DAP	Abstract		
AA:	70	10	Z=3.87, p<.0001	phi=.61
AN:	95	35	Z=3.98, p<.00005	phi=.63
NA:	75	10	Z=4.16, p<.000025	phi=.66
NN	70	10	Z=3.87, p<.0001	phi=.61
Ave:	77.5	16.25	Z=7.76, p<.0000001	phi=.61

Percentage 'P & not-Q' responses, n=20 per cell. p values are one-tailed.

Note that falsification levels are highest for AN rules. This is always the case (Evans & Lynch, 1973; Manktelow & Evans, 1979; Reich & Ruth, 1982; Yachanin & Tweney, 1982), because the "matching response" for AN rules is also the falsifying response. For a rule like "If A then not-3", the subject who chooses 'A' and '3' has both matched the values mentioned in the rule and, by coincidence, falsified.

To sum up: four SC rules were tested, and all four elicited a robust content effect, even though three of these were phrased in unconventional ways (vis-a-vis the actual Florida Drinking Age Law), and had negated components.

* for example, The "not-3" in "If A then not-3" cannot be recoded as a simple affirmative phrase.

2.4.3 The Sears Problem

The Sears problem was developed by D'Andrade (1981). It is set in a Sears store, and specifies the conditions under which a purchase must be authorized by the department manager. In real life, the purpose of such procedures is to make sure that customers or sales clerks do not cheat the store by "paying" for goods with a bad check, defunct or stolen credit cards, etc. Thus, it is a particularly nice example of a social contract problem.

D'Andrade (reported in Rumelhart & Norman, 1981)

D'Andrade's abstract and thematic problems both used prescriptive rules in the workplace. As part of your job, you (the subject) are supposed to make sure the rule has been followed. Half his subjects were given the abstract problem, half the thematic problem (this report does not say how many subjects were involved).

Both conditionals were embedded in the text of a story. In the abstract problem, "As part of your job as a label checker at Pica's Custom Label Factory, you have the job of making sure that all labels with a vowel printed on one side have an odd number printed on the other side." In the thematic problem, "As part of your job as an assistant at Sears, you have the job of checking sales receipts to make sure that any sale of over $30.00 has been approved by the section manager. (This is a rule of the store.)"

Both problems are set in the workplace, both use prescriptive rules and both invoke a "detective set" (Van Duyne, 1974) -- the subject is a "checker", looking for violations of a rule. But the Sears rule is a social contract whose violation

indicates theft -- a customer (or sales clerk) is not supposed to take something of great value without having paid its cost (as vouchsafed by the section manager). The abstract problem expresses a prescriptive rule, but not one whose terms are recognizable costs and benefits. Thirteen percent of subjects answered 'P & not-Q' for the abstract problem, whereas nearly 70% gave this answer for the social contract problem.

Griggs & Cox, 1983

Griggs & Cox's Experiment 2 included a condition identical to their DAP condition (described in 2.4.2), except that they tested systematically negated Sears problems. The abstract control groups for the Sears problems were the same as those for the DAP.

As with the DAP, Griggs & Cox systematically negated components of an AA Sears problem in a way that preserved their structures as standard social contracts. To do this, some of the negated components had to become quite long and complicated, using both explicit ("not") and implicit ("without") negatives to create what amounts to a double negative. The rules were:

AA: "If a purchase exceeds $30, then the receipt must have the signature of the department manager on the back."

AN: "If a purchase exceeds $30, then the receipt must not be without the signature of the department manager on the back."

NA: "If a purchase is not less than $30, then the receipt must have the signature of the department manager on the back."

NN: "If the [sic] purchase is not less than $30, then the receipt must not be without the signature of the department manager on the back."

For all four rule forms, the antecedent clause specifies that the rule pertains to high value purchases (large benefit), whereas the consequent clause specifies the authorization requirement

(vouchsafing that the high cost has been paid). Hence all four thematic problems are STD-SCs. The results are shown in Table 2.4.

Table 2.4 Griggs & Cox, 1983, Experiment 2 (Sears)

	Sears	Abstract		
AA:	75	10	Z=4.16, p<.000025	phi=.66
AN:	70	35	Z=2.22, p<.013	phi=.35
NA:	60	10	Z=3.31, p<.0005	phi=.52
NN:	50	10	Z=2.76, p<.003	phi=.44
Ave:	63.75	16.25	Z=6.13, p<.0000001	phi=.48

Percentage of 'P & not-Q' responses, n=20 per group.
p values are one-tailed.

Despite abstruse, double negative circumlocutions, all four SC rules elicited a sizable and significant content effect. I would guess that the smaller percentages for the negated rules is due to their increasingly complex structure of negation.

2.4.4 Deformed Social Contract Rules

Deformed social contract rules have components that are recognizable as costs and rationed benefits. However, these components are arranged such that they violate the principles of social exchange, wherein one is obliged to pay a cost in order to be entitled to a benefit. Using deformed SC rules, one can see whether any prescriptive rule involving costs and benefits can elicit a robust and reliable content effect, or whether the costs and benefits must be arranged in the format of a standard social contract.

Deformed social contracts can be generated by systematically negating components of a STD-SC. For example:

AA: "If you take the benefit, then you must pay the cost."
AN: "If you take the benefit, then you must not pay the cost."
NA: "If you do not take the benefit, then you must pay the cost."
NN: "If you do not take the benefit, then you must not pay the cost."

The AN, NA, and NN rules violate the principles of social exchange.*

The ideal way to investigate deformed social contracts is to generate them from a rule that has no counterpart in the subject's experience. This can be done by embedding an unfamiliar rule in a story that defines its terms as costs and benefits, and arranging those terms in a format that violates the principles of social exchange. Deforming an SC rule that actually exists (like the DAP) introduces unfortunate demand characteristics. Subjects may assume the experimenter is testing their knowledge of the actual rule or law.** Alternatively, using a deformed version of an actual rule may "cue" the subject to reframe it as a proper (but different) social contract in a sci-fi world, and reason accordingly. The further a cost/benefit rule is from an explicit, real-life law, the better.

Unfortunately, the only experiments on deformed social contracts use the Sears problem and the DAP (Griggs & Cox, 1983, Experiment 1). The deformed DAPs use components of the Florida Drinking-Age Law, an explicit SC rule that actually exists. They are, therefore, especially vulnerable to the demand

---

* At first glance, it might seem that, although strange, the NN rule does not exactly violate rules of social exchange. However, in social exchange, a person who has not received a benefit is never prevented from paying the cost if he still wants to (see Chapter 5). This is, however, a more minor violation than the gross assymetries posed by the AN and NA rules.

** For example, a subject who assumed an experiment with deformed DAPs is actually testing her knowledge of the Florida Drinking Age Law would choose the "drinking beer" and "16 years old" cards, regardless of which logical categories they belong to.

characteristics noted above. Deformed Sears problems are somewhat less susceptible to the same interpretational difficulties because the AA Sears problem invokes more general SC principles, rather than an explicit, existing, law.

Griggs & Cox, 1983 (Experiment 1), systematically negated the components of the (AA) Sears problem and the DAP by inserting "not" or "does not" into the antecedent and consequent terms, regardless of how this affected the rule's status as a social contract. Hence, a phrase like "then you must pay the cost" would become "then you must not pay the cost." They compared these rules to ordinary letter-number abstract rules that had been similarly negated.

Deformations of the Sears Problem

The AA Sears problem does not exist as an actual rule or law. Rather, it invokes a common method for detecting potential cheaters in situations where money is being exchanged for goods. The rules that Griggs & Cox used are listed below. A translation into cost/benefit language is listed below each rule.

AA: "If a purchase exceeds $30, then the receipt must have the signature of the department manager on the back."

"If a customer takes a high value benefit, then we must make sure he has paid the cost.

AN: "If a purchase exceeds $30, then the receipt must not have the signature of the department manager on the back."

"If a customer takes a high value benefit, then we must not make sure he has paid the cost."

NA: "If a purchase does not exceed $30, then the receipt must have the signature of the department manager on the back."

"If a customer takes a low value benefit, then we must make sure he has paid the cost."

NN: "If a purchase does not exceed $30, then the receipt must not have the signature of the department manager on the back."

"If a customer takes a low value benefit, then we must not make sure he has paid the cost."

Griggs & Cox did not think of these problems in terms of social exchange. The reason they tested these deformations was to see whether the real (AA) rule would shine through, guiding the subject to a falsifying response. On this view, the AN, NA, and NN rules should all elicit content effects, and these should be of approximately the same size.

An alternative view is that only rules that instantiate standard social contracts elicit content effects. This leads to a different set of predictions, based on the principle that the deformed rule that most violates the principles of social exchange should elicit the weakest content effect.*

Hard cases make bad law -- prediction would be easier had Griggs & Cox fabricated unfamiliar SC rules. However, social contract reasoning should elicit the following pattern of results for the AN, NA, and NN deformations of the Sears problem:

1. The NA rule violates the principles of social exchange the least. In real life, <u>all</u> purchases are benefits rationed according to ability to pay, even low cost purchases. The use of authorization signatures is a general method that can be applied no matter what the goods are, or how much they are worth: whether it is invoked for purchases which are over $30 or under $30 depends on how worried a manager is about the theft in each price category. It is certainly peculiar to require authorization of less valuable purchases,** but it does not do too much violence to the rule's status as STD-SC.

---

* Usually, no content effect at all. It depends on how easy it is to reframe a particular rule as a STD-SC.

** Especially if the NA rule leads one to infer that more valuable purchases do not require authorization. This inference is pragmatically reasonable, but logically invalid.

2. The AN rule violates the principles of social exchange the most. When a high value purchase has been made, it makes no sense to insist that a clerk refrain from making sure the cost has been paid. Furthermore, there is no easy way of reinterpreting the AN rule as a STD-SC. Thus the fact that the AN DAP suggests an SC, but is grossly deformed, should sow confusion, eliciting no content effect, or even a negative one, as compared to the AN abstract rule which so easily leads to falsification through matching.

3. The NN rule falls somewhere in between. If it said one need not check for authorization of low value purchases, rather than that one must not, it would not directly violate the principles of social exchange; in fact, it could easily be interpreted as an indirect way of stating the AA rule. However, the strange inclusion of must not might make subjects assume that both the NN and AA hold (as indicated by choosing all four cards). Pragmatic inference aside, in cost/benefit terms the NN rule has the same structure as the AN rule -- the violation is mitigated only by the fact that the antecedent of the NN rule represents a lower value benefit than that of the AN rule. Hence, it should not elicit a content effect.

To sum up: the social contract view predicts that the AN rule will elicit no content effect (or even a negative one), the NN rule will elicit no effect, and the NA rule will elicit a content effect that modest compared to that for the AA rule. The results are summarized in Table 2.5.

Table 2.5 Griggs & Cox, 1984, Experiment 1 (Sears)

	Sears	Abstract		
AA:	85	5	Z=+5.09, p<.00000025	phi=.80
AN:	40	75	Z=-2.24, p<.013	phi=.35
NA:	60	25	Z=+2.24, p<.013	phi=.35
NN:	15	15	Z= 0 n.s.	

Percentage 'P & not-Q' responses, n=20 per group. p values are one-tailed.

As usual, the AA Sears problem elicited a large, significant content effect. The rest of the problems follow the social contract predictions. The NA Sears problem, which is a peculiar, but still proper, STD-SC, elicited a more modest content effect

than the AA rule. The AN Sears problem, which grossly violates the principles of social exchange, actually elicited a negative content effect -- that is, subjects did better on the AN abstract problem than they did on the AN Sears problem. The NN Sears problem did not produce a content effect. Furthermore, more subjects chose all four cards on the NN Sears problem than on the NN abstract problem (40% v. 10%: $Z=2.19$, $p<.015$, phi=.35). This is what one would expect if subjects made the pragmatic inference that the NN Sears rule also implies the STD-SC AA Sears rule, but did not make the equivalent inference on the NN abstract rule.

This pattern of results is not predicted by Griggs & Cox's view -- in fact, they found the results puzzling. On their view, the AN and NN rules should have elicited content effects, just as the NA rule did. Griggs & Cox thought the negated rules would bring the AA Sears rule to mind and guide their inferences through "reasoning by analogy" to the logically falsifying choice. If subjects engaged in any reasoning by analogy to the AA rule, then the analogy must have been to the AA rule's structure as a standard social contract rather than to its structure as a logical conditional, because results for negated rules are best predicted by their cost-benefit structure.

Deformations of the DAP

The DAP does not invoke a general procedure for detecting potential cheaters, as does the Sears problem. Rather, it is a straightforward version of the Florida Drinking-Age Law, a specific, explicit law that was quite familiar to Griggs & Cox's University of Florida subjects. Thus, one can expect subjects' interpretations of deformed DAPs to be less flexible and more

vulnerable to demand characteristics than their interpretations of deformed Sears problems. The DAP rules used, and their translations into cost/benefit language, are listed below.

AA: "If a person is drinking beer, then the person must be over 19."
"If you take the benefit, then you must pay the cost."

AN: "If a person is drinking beer, then the person must not be over 19"
"If you take the benefit, then you must not pay the cost"

NA: "If a person is not drinking beer, then the person must be over 19"
"If you do not take the benefit, then you must pay the cost"

NN: "If a person is not drinking beer, then the person must not be over 1
"If you do not take the benefit, then you must not pay the cost"

Griggs & Cox's view makes the same prediction as for the deformed Sears problems: the AN, NA, and NN rules should all elicit moderate content effects. As before, the social contract view has a different set of predictions:

1. The AN DAP severely violates the principles of social exchange. It should elicit no content effect, or even a negative content effect, for the same reasons that the AN Sears problem should.

2. The NA DAP is not so close to the AA DAP as the NA Sears problem is to its AA counterpart. Despite the negation, the antecedent of the NA Sears problem refers to a benefit that is rationed by ability to pay (purchases under $30), just as the antecedent of the AA Sears problem does. However, the antecedent of the AA DAP refers to a benefit that is rationed by age ("drinking beer"), whereas the antecedent of the NA DAP refers to a benefit that is not rationed by age ("drinking coke"). The NA Sears problem was merely a weak, or peculiar, STD-SC, whereas the NA DAP actually violates the principles of social exchange. Hence the NA DAP should not elicit a content effect.

   However, there are reasons to believe the NA DAP will sow less confusion than the AN DAP. "Drinking coke" -- the NA DAP's negated component -- is not generally considered a liability by persons of any age. Thus, although "drinking coke" is not an age-rationed benefit, it is, at least, a benefit. In contrast, "must not be over 19" -- the AN DAP's negated component -- is not a cost/requirement for anything in our society. I can think of no benefit in our society that is open to adolescents but not to adults. Hence, the NA DAP violates the principles of social exchange somewhat less than the AN DAP. It is therefore less likely to elicit a negative

content effect than the AN DAP.*

3. Like the NN Sears problem, the NN DAP suggests a host of pragmatic inferences, which makes prediction difficult. The AA DAP is so well known that subjects undoubtedly realize it does not imply the NN DAP. Thus, it is unlikely that anyone would think that both the NN DAP and AA DAP hold, and choose all four cards. The NN DAP is certainly counter to experience, in that "drinking coke" is not a rationed benefit in our society, and "must not be over 19" is not a cost/requirement for anything. It may be so clearly counter to experience that it invites reframing as a proper STD-SC in a sci-fi world -- a world where adolescents take their revenge, and adults are not permited the benefit of drinking coke, the drink of the adolescent. After all, the subjects were college students who were, or recently had been, below the legal drinking age. If some subjects made this sci-fi conversion, the NN DAP could elicit a modest content effect. This is why it is better to use unfamiliar rules -- existing laws invite too many interpretations to make sound prediction possible. Hard cases make bad law.

To sum up: The social contract view predicts no content effect (or a negative one) for the AN DAP, no effect for the NA DAP, and a question mark for the NN DAP.

As Table 2.6 shows, the results do not conform to Griggs & Cox's prediction of a content effect for all four problems -- the AN and NA DAPs did not elicit content effects at all. The pattern of results is best predicted by the social contract view.

The AA DAP, a STD-SC, elicited its usual hearty content effect. The AN and NA DAPs, which violate the principles of social exchange, did not elicit a content effect. Moreover, the AN DAP -- like the AN Sears problem -- was the only thematic

* The results of Evans & Lynch, 1973, also suggest that the NA rule might sow less confusion. Choice of the 'P' card is least influenced by matching. 'P' represents a true antecedent, and it is almost universally chosen due to a rudimentary understanding of logic or contingency. Thus, when a subject needs to reframe a bizzare rule to make sense of it, one might expect the subject to show greatest flexibility in re-interpreting the antecedent, and least flexibility in re-interpreting the consequent. If so, re-interpreting the AN DAP's negated consequent might be more difficult than re-interpreting the NA DAP's negated antecedent.

Table 2.6 Griggs & Cox, 1983, Exp 1 (DAP)*

	DAP	Abstract			DA Law
AA:	70	5	Z=+4.25, p<.000025	phi=.67	--
AN:	55	75	Z=-1.36, n.s		15
NA:	30	25	Z=+0.35, n.s.		30
NN:	40	15	Z=+1.77, p<.04	phi=.28	25

Columns 1 and 2: Percentage 'P & not-Q' responses, n=20 per group. The "DA Law" column shows the percentage of subjects who chose cards that indicate their knowledge of the actual Florida Drinking Age Law, irrespective of logical category. p values are one-tailed.

problem to elicit fewer falsifying responses than its abstract counterpart. Although the 20% difference between the AN DAP and its abstract counterpart falls 6 points short of a significant negative content effect, only the social contract view predicts that the AN data will move in this direction.

The NN DAP elicited a small content effect (phi=.28). This result neither confirms nor denies the social contract view, which makes no strong predictions about the NN DAP. However, there is evidence that knowledge of the actual drinking-age law prevented subjects from assuming that both the AA and NN DAPs hold: only 5% of subjects chose all four cards on the NN DAP, compared to 40% of subjects on the NN Sears problem. This difference is significant (Z=2.65, p<.005, phi=.42).

Knowledge of the Florida law appears to have introduced other demand characteristics. The DA Law column of Table 2.6 shows the percentage of subjects who chose "drinking beer" and "16 years old", regardless of which logical categories these cards belonged to. These are the cards one would choose if looking for cheaters on the real drinking-age law. Because the negated DAP rules have quite obviously been yanked from an

existing SC law, subjects given a deformed DAP are far more likely to think the experimenter is testing their knowledge of the actual law than subjects given a deformed Sears problem.* The evidence supports this view. Significantly more subjects chose cards that would falsify the corresponding AA, STD-SC rule for negated DAPs than for negated Sears problems (23% v. 8%: $Z=2.25$, $p<.013$, phi=.21).

The results of Griggs & Cox's two experiments with deformed social contracts indicate that rules that violate the principles of social exchange do not elicit content effects. The closer a rule comes to the format of a standard social contract, the more likely it is to elicit a thematic content effect.

Social Contract Summary

Sixteen out of sixteen experiments comparing social contract rules to abstract rules have produced a robust content effect (Johnson-Laird et al., 1972; Van Duyne, 1976; D'Andrade, 1981; Golding, 1981; Griggs & Cox, 1982; Cox & Griggs, 1982; Griggs & Cox, 1983 (10 replications)). When a prescriptive post office problem was administered to cultural groups for whom the constituents were not recognizable as costs and rationed benefits -- that is, when subjects did not perceive the rule as a social contract -- no content effect was found (Golding, 1980; Griggs & Cox, 1982). Deformed social contracts -- rules that share constituents with proper social contracts but grossly violate the principles of social exchange -- do not elicit content effects (Griggs & Cox, 1983).

---

* only 1-2 subjects per negated Sears problem gave answers consistent with this interpretatation of the experiment.

General Summary of Chapter 2

Robust and replicable content effects are found only for rules that relate perceived benefits to perceived costs in the format of a standard social contract. No thematic rule that is not a social contract has ever produced a content effect that is both robust and replicable. For thematic content areas that do not express social contracts, either no content effect is found (the food problem), or there are at least as many studies that do not find content effects as there are studies that do (transportation and school problems). Moreover, most of the content effects reported for non-SC rules are either weak (Gilhooly & Falconer, 1974; Pollard, 1981), clouded by procedural difficulties (Bracewell & Hidi, 1974; Van Duyne, 1974), or have some earmarks of a social contract problem (Van Duyne, 1974). All told, for non-SC thematic problems, three experiments have produced a substantial content effect (transportation: Wason & Shapiro, 1971; Bracewell & Hidi, 1974; school: Van Duyne, 1974), two have produced a weak content effect (transportation: Gilhooly & Falconer, 1974; Pollard, 1981), and 14 have produced no content effect at all (transportation: Bracewell & Hidi, 1974; Manktelow & Evans, 1979; Yachanin & Tweney, 1982; Griggs & Cox, 1982. food: Manktelow & Evans, 1979 (4 experiments); Brown et al., 1982; Reich & Ruth, 1982; Yachanin & Tweney, 1982; school: Yachanin & Tweney, 1982. non-SC post office: Golding, 1980; Griggs & Cox, 1982). The few effects that were found did not replicate. In contrast, sixteen out of sixteen experiments with standard social contracts elicited substantial content effects. These include the Drinking Age Problem, the Post Office Problem, and the Sears

Problem. Deformed social contracts, which share constituents with standard social contracts but grossly violate the principles of social exchange, do not elicit content effects.

In this extensive literature, standard social contract rules are the only thematic content rules to elicit strong, replicable content effects on the Wason selection task.

Chapter 3

"Differences in Experience": Proposed explanations for the elusivity of the content effect on the Wason selection task

A number of theories attempting to explain the elusive content effect on the Wason selection task have appeared in the literature. Most agree that thematic content enhances logical performance because thematic rules are familiar, whereas abstract rules are unfamiliar. The theories differ in their explanations of why familiarity enhances performance, and why the content effect is so "elusive."

None of these theories invoke the notion of a social contract, therefore none of them try to explain why social contract rules are the only thematic rules to consistently elicit robust content effects. Invoking the concept of a social contract turns the theoretical problem on its head: the phenomenon requiring explanation is not the content effect's elusivity, but, rather, its predictability.

3.1 Families of explanation

Before discussing the particular theories that have already been proposed, it is useful to consider what kinds of explanation are possible <u>in general</u>. Conceptually, there are at least five relatively distinct families of explanation:

1. There is no logic module. In solving the selection task, people use rules of inference appropriate to the domain suggested by the problem. These rules of inference may be different for different content domains.

a. The rules of inference are a product of "experience" structured only by information processing mechanisms that are innate, but domain general.

b. The rules of inference are innate, or else the product of "experience" structured by domain specific innate algorithms.

2. There is a logic module, but it is not necessary for everyday learning. It is activated only in higher level model building, for example, to answer questions within the framework of a well-established theory of what is true of a content domain. That is why performance is better with familiar materials.

3. There is a logic module, and it is necessary for learning. The content effect is due to differences in how well the propositions can be pushed through auxiliary mechanisms like short term stores or imagery buffers. Familiar terms and/or relations facilitate performance because they are concrete and therefore more easily manipulable or because they reduce "cognitive load".

4. There is no logic module, just the ability to recognize contradiction when one sees it.

   a. People can build mental models of the circumstances described in a problem; if they happen to build a model that contradicts the state of affairs asserted by the conditional, they will falsify. It is easier to build mental models of familiar propositions and relations.

   b. Actual experience with events that contradict the relation are stored in long-term memory. A familiar theme is more likely to cue contradictory associational pairing from long term memory, because such pairings are more likely to have been actually experienced.

5. Non-rational, domain-general heuristics having nothing to do with formal logic, or with an understanding of the relevance of counter-examples, account for the presence and variability of the content effect.

The hypothesis that humans have Darwinian algorithms for reasoning about social exchange is a "family 1-b" explanation. Each explanation proposed in the literature belongs to one of these five families of explanation.

3.2 Explanations proposed in the literature

A number of explanations have been put forth to explain content effects on the Wason selection task. Most of them involve a wedding of associationism and Tversky & Kahneman's (1973) "availability" heuristic.

Tversky & Kahneman were interested in how people judge probability. They noted that people typically do not make statistically sound probabilistic inferences, even when given information sufficient to do so.

Although people's probability judgments are not statistically sound, they are not random, either. To account for this, Tversky & Kahneman posited that people use mental short cuts -- "heuristics" -- in making probability judgments. They hypothesized that people judge the probability of two events co-occurring by the ease with which examples come to mind -- their "availability". They named this method the "availability heuristic."

For example, suppose you are told that 80% of college students in Cambridge attend Harvard and 20% attend MIT. A Cambridge college student was involved in a fight today. Your task is to guess which school this student attends. Five fights involving MIT students immediately spring to mind, but you have to search your memory long and hard to recall any fights involving Harvard students: the co-occurrence of "MIT" and "fight" is more available as a response. Even though Harvard students outnumber MIT students 4 to 1 in Cambridge, and even though you have no reliable data indicating that MIT students are more pugnacious than Harvard students, the availability heuristic

would lead you to judge that the fight today was more likely to have involved an MIT student than a Harvard student.

According to Tversky & Kahneman, ease of recall is a function of associative strength. Associative strength, they argue, is usually directly proportional to the frequency with which two events co-occur in an individual's experience. The availability heuristic is a useful rule of thumb because the ease with which associations can be brought to mind is usually correlated with their ecological frequency. It can lead to bias, however, when associative strength is determined by factors other than ecological frequency (like semantic distance or perceptual saliency).

Frequent events are familiar events. Abstract rules relating letters and numbers are unfamiliar. It occurred to a number of researchers that availability -- based on frequency-determined associative strength -- might play a key role in explaining why some familiar problems are more likely to elicit logical performance on the Wason selection task than abstract problems.

For Tversky & Kahneman, ecological frequency was only one of many determinants of availability. But because selection task theorists were trying to account for a content effect that they assumed was caused by familiarity, associationism plays a more central role in their adaptations of availability theory.

The "availability theories" of the selection task theorists come in a variety of forms, with some important theoretical differences. But common to all is the notion that the subject's actual past experiences create associational links between terms

mentioned in the selection task. The more exposures a subject has had to, for example, the co-occurrence of P and Q, the stronger that association will be and the easier it will come to mind -- become "available" as a response. A subject is more likely to have actually experienced the co-occurrence of P and not-Q for a familiar rule, therefore familiar rules are more likely to elicit logically falsifying responses than unfamiliar rules. If all the terms in a task are unfamiliar, the only associational link available will be that created between P and Q by the conditional rule itself, because no previous link will exist among any of the terms. Thus 'P & Q' will be the most common response for unfamiliar rules.

Although it is rarely explicitly stated, these theorists seem to assume that associative links are created "the old-fashioned way", by domain general associative processes. Some refer directly to associationism (Pollard, 1982), whereas others refer more simply to the different amount of "experience" subjects may have had with different content domains (Griggs & Cox, 1982; Manktelow & Evans, 1979; Johnson-Laird, 1983; Wason, 1983). The presumption that learning occurs via some sort of "computational associationism" (Fodor, 1983) would account for their belief that the categorization of content domains along a familiar-unfamiliar dimension is the correct one, the one with causal import. Associationism is a process that makes unfamiliar content domains familiar -- regardless of the specific content of the domain it operates upon. Which content domains become familiar is determined by the amount of personal experience a particular individual has with the domains in question. The

selection task theorists rarely entertain the notion that regardless of familiarity, different content domains are processed by different, domain specific rules of inference. When they do, they seem to presume that the domain specific rules were learned via a domain general process.

The P card is almost universally chosen on Wason selection tasks, regardless of content. All theories that have been proposed in the literature concede that this is probably due to a rudimentary understanding of logic (or of contingency, in a looser, linguistic sense). Thus, the primary goal of these theories is to explain why familiar rules facilitate the selection of the not-Q card and inhibit the selection of the Q card, insofar as this happens. To be adequate, a theory must be able to answer three questions raised by the data reviewed in the previous chapter:

1. Why do familiar rules elicit more logical falsification than abstract rules?
2. Why do some familiar rules reliably elicit logical falsification whereas others do not?
3. Why do the same familiar rules sometimes elicit logical responses and sometimes not?

Differential Availability

In an article entitled "Human reasoning: Some possible effects of availability", Paul Pollard put forth what is to date the most precisely specified theory purporting to explain content effects on the Wason selection task (Pollard, 1982). It is a quite literal application of the associationist paradigm sketched above, in which pre-existing associative pairings of terms

mentioned in the selection task create a non-logical response bias (his theory is a "family 5" explanation). Whether a subject responds 'P & Q' or 'P & not-Q' is determined by the relative strength of these two associative links.* The dominant association wins, even if both are available. Thus, a subject will answer 'P & Q' if more instances of P - Q links come to mind than instances of P - not-Q links. For Pollard, associative strength is directly proportional to the number of exposures an individual has had to each pairing. Actual personal experience is the centerpiece of his availability theory.

For example, on a transportation problem where the rule is "If a person goes to Boston then he takes the subway" and the cards are "Boston" (P), "Arlington" (not-P), "subway" (Q), and "cab" (not-Q), a subject who had had more experiences of people taking the subway to Boston would choose "Boston" and "subway", that is, 'P & Q'. A subject who had had more experiences of people taking a cab to Boston would choose "Boston" and "cab", that is, 'P & not-Q', which is, by coincidence, the logically falsifying response. Note, however, that the procedure that generated this response is decidedly non-logical.

Pollard distinguishes between "realistic" content and content that is merely "thematic". Thematic content is not

---

* Pollard does not explicitly discuss why someone might choose 'P' alone on the selection task. However, in discussing other logical tasks he notes that availability might affect a conditional's perceived reversibility; "all dogs are animals" is clearly not the same as "all animals are dogs", whereas "all dogs bark" is not so clearly different from "all barking animals are dogs." From here he would have to argue that having understood that "If P then Q" does not imply "If Q then P" somehow prevents one from choosing the Q card. But since his theory is a nonlogical one, and nonreversibility is a logical consideration, it is not clear what that "somehow" would be.

"realistic" unless it cues actual experiences. If the subject has had no relevant experiences with the problem domain, no matter how "thematic" it is, the dominant association will be that created by the conditional rule itself. Hence, the subject will respond 'P & Q', just as if the problem's content were abstract. Pollard is a stickler for actual experience. For example, I can think of no relation that people have more experience with than that expressed by the food problem: "If I eat X then I drink Y". Most meals include both food and drink, and most people eat three such meals a day, every day of their lives. Moreover, it is quite common for certain foods and drinks to be consumed in conjunction with one another: orange juice with breakfast foods, coffee with dessert, wine with dinner entrees, mixed drinks with hors d'oeuvres.* Yet Pollard claims that the food problem did not elicit a content effect because subjects probably had not personally experienced some of the particular food-drink combinations used, such as, "If I eat haddock then I drink gin". (In some of my experiments I administered food problems using more usual content, and still found no effect, see Chapter 6.)

Because responses are determined by the actual, personal, idiosyncratic experiences of subjects, his theory can account for the fact that certain contents, like the transportation problem, sometimes elicit logical responses and sometimes do not:

---

* Note also that for meals, the most common eating plus drinking experiences, it is the type of food eaten that determines what drink is served, not vice versa. Like the food problem, the relation for meals is "If I eat X, then I drink Y", not "If I drink Y then I eat X).

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

Unfortunately, for the same reason, his theory has very little predictive power. For a particular subject population, one can generate predictions if the problem's content taps experiences that the experimenter knows to be nearly universal or else completely unfamiliar. But for most content domains, the only prediction it can make is that responses will vary unpredictably.

The fact that the Drinking Age Problem (DAP) and Johnson-Laird et al.'s post office problem elicit high percentages of 'P & not-Q' responses presents difficulties for Pollard's theory. Most subjects have had more exposures to beer drinkers who are over 20 (legal) than under 20 (illegal) and seen more envelopes with correct postage than with incorrect postage. Thus, an implication of his differential availability view is that most subjects will choose 'P & Q' for these problems. Pollard notes this difficulty and tries to finesse it by suggesting that differential availability arises from the subject's experience of the content <u>and</u> <u>context</u> of the problem. He says:

> The context relates to drinkers that are investigated by the police, or drinkers who are breaking the law, and the only available instances of these, given the context, are underage drinkers (or, in the case of the Johnson-Laird et al. study, understamped letters). The P - not-Q link thus becomes dominant. (p. 80)

This explanation is problematic. Unless you already understand that "breaking the law" = P + not-Q, playing the role of a police

officer or postal sorter seeking violations of the law will not, in and of itself, limit your search to instances of not-Q (underaged beverage drinkers, understamped letters). This criticism is underlined by results on the post office problem for Golding's younger subjects and Griggs & Cox's American subjects. These subjects did not understand that "violating the rule" = sealed envelope + less than 20 cents postage. Playing the role of a postal sorter looking for violations did not help them one wit, even though this is the same context successfully used by Johnson-Laird et al. To look for a violation you have to know what counts as a violation. And if you already know what counts as a violation, then why not answer the selection task accordingly? Why would the relative availability of compliance versus violation episodes cause you to change your answer?

One could reframe Pollard's view of context thus: Most subjects have had experience with the police and have noted that they only question people under 20, and this makes not-Q more available than Q. But is this true? Police do not investigate guilty people only -- they query a range of people in search of the guilty. In my experience, bouncers (I have never witnessed police making such inquiries) ask to see the IDs of people who look young -- but most of these prove to be over the legal drinking age. I suspect my experience is not atypical. So all people sharing my experience of bouncers/police should choose 'Q' rather than 'not-Q'. And how many people have had any actual experience with postal sorters, to see what sorts of envelopes they pay special attention to? The point is, subjects' experience with the behavior of police and postal sorters is

bound to be as idiosyncratic as their experience with going to Boston via cab or subway. Therefore, if we reframe Pollard's view of context in this manner, responses to the DAP and post office problems should be variable, like those to the transportation problem. They should not elicit such uniformly high levels of falsification.

Last, Pollard seems to pick and choose that which he wishes to count as "actual experience." The subject, who has never been, and perhaps never even met, a postal sorter, can project himself into this role such that this imagined person's long term memory is cued. Yet this same subject cannot make the intuitive leap from haddock with water to haddock with gin. I can see no principled way of maintaining that the transportation problem and post office problem cue familiar experiences, but that the food problem does not.

Memory-cueing/ Reasoning by analogy

Memory cueing (Manktelow & Evans, 1979; Griggs & Cox, 1982; Cox & Griggs, 1982; Griggs, 1983) is a variety of availability theory that does not depend on differences in the relative availability of P & Q versus P & not-Q. Although it was first suggested by Manktelow & Evans (1979) to explain why the thematic content effect is so elusive, its primary proponents are Richard Griggs and James Cox (Griggs & Cox, 1982; Cox & Griggs, 1982; Griggs, 1983). It is a "family 4-b" explanation.

According to these researchers, subjects will falsify on the Wason selection task if they can recall past experience with:

1. the content of the problem;
2. the relationship (rule) expressed; and
3. a counter-example to the rule.

Recalling past experience with all three aspects of the problem allows the correct response to be "cued" from long term memory.

Unlike Pollard's differential availability theory, which requires that available disconfirming instances outnumber available confirming instances, memory-cueing theorists only require that one counter-example become available. Subjects do not* generate falsifying instances by a deductive process, but if a counter-example happens to be generated by some other means, they can recognize it as violating the rule. This highlights an important conceptual difference between differential availability theory and memory-cueing. Differential availability is an entirely nonlogical theory, whereas memory-cueing requires minimal logical competence: the ability to recognize contradiction, the most fundamental logical property.

The experiments reported by Griggs & Cox, 1982, was very important in establishing memory-cueing as a theory. The transportation and post office problems failed to elicit more logical responses from their American subjects than abstract problems did. However, 72% and 74% of subjects from the same population produced falsifying responses in two different replications of the DAP. Griggs & Cox substantiated their claim that members of their subject pool had past experience with the above three aspects of the DAP, but not with the post office problem. Thirty-three additional subjects from the same population completed a questionnaire designed to tap their

* and cannot, without explicit academic training in formal logic.

familiarity with these two rules and counter-examples to them. The questionnaire asked:

1. whether regulations exist concerning beer and being of a certain age, and sealing an envelope and having a certain amount of postage on it; if so, then write the regulation,
2. whether they themselves had ever violated the regulation,
3. whether they could remember specific instances of someone other than themselves violating the regulation.

Only 12% wrote a rule consistent with the post office problem, but 88% wrote a rule consistent with the DAP. Only one subject recalled having personally violated the postal rule (interesting, as no such rule exists in the U.S.!). In contrast, 76% of subjects reported having personally violated the drinking age rule, and 97% could recall specific instances of someone else violating it.

Griggs & Cox take this correlation of personal experience in their subject population with success on the selection task as evidence for memory-cueing theory. They also cite Golding (1981), in which older subjects who were familiar with Britain's pre-1968 post office rule did well on the post office problem, whereas younger subjects did not.

They explain the inconsistency of the results for other thematic problems (food, schools, transportation) as caused by the variable, idiosyncratic, experience of subjects with these contents. They suggest, for example, that Wason & Shapiro's (1971) transportation problem elicited higher levels of logical falsification than those of Manktelow & Evans (1979) and Pollard (1981), because Wason & Shapiro's subjects from University College London live closer to the cities named in the selection task than do the other researchers' Plymouth Polytechnic

subjects. Therefore, Wason & Shapiro's subjects were more likely to have made a trip that happened to be a counter-example to the rule.*

Note that the fact that Griggs & Cox hazard this explanation for the transportation problem means that they only require that the subject have experience with the relation expressed by the rule. Subjects needn't have experienced the rule qua rule -- that is, as an explicit, linguistically expressed set of propositions, such as the DAP and the British postal office rule.

If memory-cueing is the full story, one wonders why performance on food problems is so uniformly low. Although memory-cueing requires that the subject have had experience with a counter-example, it does not require that the subject have had experience with the exact counter-example suggested by the uncovered not-Q card. On the DAP, for example, the subject can still be expected to choose a not-Q card that says "16 years old" even if her specific experience was of an 18 year old illegally drinking beer. The food problem studies do not report what food and drink pairs they actually used, but some authors (e.g. Pollard, 1981) have made of the fact that Manktelow & Evans' instructions used some rather odd combinations, such as, "If I eat haddock, then I drink gin". But the odder the combination, the higher the probability that a subject would have experienced

* This explanation would have difficulty accounting for Bracewell & Hidi, 1974: Even though both transportation problems were given to the same subject population, one linguistic format elicited a content effect, but the other did not. However, Griggs (1983) considers Bracewell & Hidi's instructions regarding non-reversibility too serious a confound to merit an explanation of this inconsistency.

a counter-example -- it may be true that not many people drink gin with their haddock, but I'll wager a great many have washed it down with water. The average 20 year old subject who eats three meals a day will have experienced almost 22,000 eating plus drinking events. Whatever the rules actually were, one would expect that 22,000 separate experiences would be sufficient to trigger a good number of counter-examples -- especially if many of the rules expressed odd combinations. Shouldn't the memory-cueing theorist expect a relatively consistent content effect for the food problem?

How does memory-cueing theory handle D'Andrade's Sears problem? As Griggs (1983) notes, chances are that very few subjects have been assistants to Sears' managers, or even worked in a store that required managers to authorize receipts. To handle such cases, Griggs and Cox couple "reasoning by analogy" with memory-cueing theory. Griggs (1983) points out that Johnson-Laird et al.'s British subjects did just as well on the post office problem when the stamps were Italian rather than British. He argues that this is because the familiar rule using pence is <u>analogous</u> to the unfamiliar rule using lire. He explains D'Andrade by saying that most subjects have probably had experience with analogous situations, such as store managers authorizing the subject's own check. "What seems to be essential is that the problem cue the subjects to recall their experience with the specific situation <u>or</u> analogous situations" (Griggs, 1983, p.26).

Cox & Griggs (1982) argue that they have found further support for reasoning by analogy in some experiments on transfer.

They created an "Apparel Color Problem" (ACP) which is identical to the DAP, except that the rule for the "police officer" to enforce is: "If a person is wearing blue, then the person must be over 19." Obviously, no subject has ever experienced such a rule. They gave each subject three problems to solve: an abstract problem (AP), the ACP, and the DAP. Cox & Griggs demonstrated that significantly more subjects solve the ACP when it comes after the DAP than when it comes before the DAP (75% v. 25%). Their explanation was that when the ACP followed the DAP, subjects reasoned by analogy to the DAP.

Interestingly, the ACP elicited a small but significant thematic content effect even when it preceded the DAP (ACP: 25%, AP: 4%). Griggs (1983) asserts that although the ACP does not relate directly to subjects' experience, they would have been in many natural situations in which their age constrained what they could do: drinking alcohol, driving, voting. Thus, the ACP could have cued one of these rules for some of the subjects, who could then "reason by analogy."*

Unfortunately, grafting reasoning by analogy onto memory-cueing theory transforms it from a moderately specified theory into an unspecified theory. What dimensions of a situation are psychologically real for subjects? Which are most important in judging similarity? How many characteristics must be shared before a subject decides that two problems or situations are

* Cox & Griggs present other data which they also interpret as instances of reasoning by analogy, using permutations of the DAP, like "If a person is over 19 then he must be drinking beer" and "If a person is under 19 then he must be drinking coke". However, these experiments are so fraught with demand characteristics of the kind described for deformed social contracts in Chapter 2 that they are difficult to interpret.

"analogous"? These are key questions, yet they are never addressed by advocates of reasoning by analogy. Without answers to questions like these, memory-cueing/reasoning by analogy explanations are ad-hoc. In the absence of a theory of analogy, reasoning by analogy guts memory-cueing theory of its predictive value.

This can be seen by considering some possible theories of analogy. For example, are the DAP and ACP are analogous because they share the same consequent term? Apparently this is not a necessary condition, because Johnson-Laird et al.'s post office problems used different terms: 50 lire stamps versus 5 pence stamps.

But perhaps problems are analogous when their consequents belong to the same class of entities* -- after all, 50 lire stamps and 5 pence stamps are still stamps. If this is the case, then why is performance so poor on food problems? There are natural situations involving explicit food rules ("If I eat red meat, then I drink red wine"; "If I eat fish, then I drink white wine"), and many involving implicit rules and relations ("If I eat breakfast cereal, then I drink orange juice", "If I eat hot chili peppers, then I drink water", "If I eat caviar, then I drink champagne", "If I eat Chinese food, then I drink tea"). These rules differ from the ones subjects were given only in the <u>particular</u> foods and drinks mentioned, just as the postal rules rrdiffered only in the particular types of stamps.

---

* Of course this begs the question. One still would need to know what dimensions are salient for deciding whether two entities belong to the same category. This formulation merely pushes the problem back one step.

The memory-cueing theorist cannot explain this difference away by pointing out that the British postal rule was explicitly mentioned in the task, for two reasons. First, this was not always true -- some subjects encountered the lire rule before the pence rule, and did very well, nonetheless. Second, Griggs (1983) attributes success on the Sears problem to "memory-cueing of general experience" (p. 25). If such general experience can be cued for check authorization, then surely it can be cued for the food problem. The same goes for the transportation problem. Isn't it likely that most subjects have favorite -- even exclusive -- ways of traveling to certain places? They walk to classes, they fly home at Christmas, they see their parents drive to work every day. Why can't they use these commonplace experiences to "reason by analogy" on the transportation problem? As mentioned above, Griggs & Cox require experience only with the relation, not with an explicit, rule. Unfortunately, Griggs and Cox never confront these questions.

It is quite possible, even likely, that people reason by analogy. It is even possible that this technique is only effective when combined with memory-cueing. My point is, until psychologists start developing theories of analogy, this variant of memory-cueing theory lacks any empirical content.

Mental Models

The mental models approach was developed by Philip Johnson-Laird (Johnson-Laird, 1982; Johnson-Laird, 1983). Explaining content effects on the Wason selection task was not his primary motivation in developing this theory. Insofar as it does account

for content effects, it relies on a form of availability. I include it because it represents a quite different view of how humans reason than do the theories previously described. Mental models is a "family 4-a" explanation.

According to Johnson-Laird, the human mind has no computational procedures that correspond to rules of inference (like modus ponens or modus tollens). Instead,

1. reasoners interpret premises by constructing an integrated mental model of them in working memory, and
2. reasoners have one piece of semantic information: A conclusion is true if the premises are true and there is no way of interpreting them so as to render it false.

These two factors can lead to logical reasoning. For example, given the premises, "Some of the scientists are parents" and "All the parents are drivers", the subject will first construct a mental model of the relation expressed by the first premise, perhaps like this:

```
scientist
scientist = parent
scientist = parent
            parent
```

The first person is a scientist who is not a parent, the second and third are scientists who are parents, the fourth is a parent who is not a scientist. All four possibilities are consistent with the premise "Some of the scientists are parents." Next, the subject will try to integrate the information in the second premise into the model of the first premise:

```
scientist
scientist = parent = driver
scientist = parent = driver
            parent = driver
```

This integrated mental model is consistent with two tentative conclusions: "Some of the scientists are drivers" (a valid

inference) and "Some of the scientists are not drivers" (an invalid inference). But which one will the subject choose? This is where the second factor enters the picture. According to Johnson-Laird, people know that a conclusion is true when the premises are true and there is no way of interpreting the premises so as to render it false. Therefore, the subject will search for alternative mental models that are also consistent with the premises, to see if any violate a tentative conclusion they have drawn. For example, the following two models are also consistent with the premises:

```
scientist =          driver          scientist = parent = driver
scientist = parent = driver    and   scientist = parent = driver
scientist = parent = driver                      parent = driver
            parent = driver
```

However, both render false the conclusion "Some of the scientists are not drivers." In contrast, both models are consistent with the conclusion "Some of the scientists are drivers."

Thus, mental modeling theory is very different from memory-cueing theory. According to memory cueing theory, people can recognize a counter-example as such if they happen to recall one, but they do not actively search for counter-examples. Also, in memory cueing theory people do not model the premises -- the premises function primarily as recall cues.

Johnson-Laird (1983) integrates content effects into his theory thus:

> If subjects already possess a mental model of the relation expressed in the general rule, or a model that can be readily related to the rule, they are much more likely to have an insight into the task. (p. 33)

He believes that memory is important in that "no effect of content can be explained without appeal to previous experience."

Previous experience gives one a library of mental models. Realistic content makes mental models available, not mere associations.

Johnson-Laird makes no attempt to predict what kinds of content will enhance performance beyond saying that familiarity with the rule helps. However, the subject need not have experienced an explicit rule (like the DAP); he cites Wason & Shapiro's original transportation problem, D'Andrade's Sears problem, and his own lire version of the post office problem as examples. However, he provides no explanations regarding why the food problem never enhances performance, why results with the transportation and school problems are so spotty, or why results with what I have called "social contract problems" are so consistent.

Frames and Schemas

At present, explanations of content effects on the Wason selection task in terms of frames or schemas are promising, but meta-theoretical. Wason & Shapiro (1971), Wason (1983), and Rumelhart & Norman (1981) have argued that reasoning on the Wason selection task is guided by frames or schemas -- domain specific inference procedures and/or mental models. These develop content area by content area, according to the subject's personal experience. The more experience a person has had with a given content area, the more likely it is that she has acquired a frame governing inference in that area. The presumption seems to be that the processes underlying the acquisition of frames are domain general, making this a "family 1-a" explanation. However,

this view would not be compromised if most frames were built by domain specific algorithms.

Although this view is akin to Johnson-Laird's mental modeling theory, it is more inclusive. Schemas or frames can enhance performance by virtue of their ability to unite the terms of the selection task into one, unified mental representation that can be easily manipulated via the frame's procedures (Wason & Green (1984) present some evidence for this view using a very simple "reduced array selection task", or RAST).* Alternatively, performance can be enhanced via the domain specific inference procedures that the schemas or frames embody.

The inference procedures that develop in a given content domain need not be logical in character. In Johnson-Laird's theory, the subject's knowledge that counter-examples are relevant to the logical validity of a conclusion is an important factor in rejecting tentative conclusions. In frame theory, the subject could be judging the soundness of a conclusion using "pieces of semantic information" that have nothing to do with logical validity. For example, the subject's knowledge of the social factors governing commercial transactions might guide her response to D'Andrade's Sears problem. This knowledge can be either declarative or procedural. Because each content area may

* The RAST is a selection task which uses only Q and not-Q cards, and usually many instances of each. Given the rule "All triangles are white", the subject is to determine whether it is true by asking to inspect -- one at a time -- the minimum number of black shapes or white shapes. The best answer is to choose all and only the black shapes; however, one can test varying degrees of insight into different rules by seeing if subjects choose more confirming white shapes for one rule than for another. The RAST is different enough from the full selection task that results on it are not directly comparable.

have different rules of inference associated with it, a frame need not lead to a logically correct answer.

The frame theorists have not yet addressed questions like: Are some content areas more likely than others to have frames associated with them? How many experiences with a domain must one have to develop a frame? Must those experiences be of a particular kind? How does the mind parse the world into separate domains for the purpose of building frames? To what extent do different individuals share the same frames?

Without answers to questions like these, the frames explanation cannot be evaluated by appeals to empirical evidence. In principle, any content effect or non-effect is compatible with it. If a particular content elicits an effect, that is post-hoc evidence for the existence of a frame for that content domain. If it does not, that is post-hoc evidence for the lack of a frame for that content domain. The Wason selection task may indeed turn out to be a useful tool for <u>discovering</u> what sort of frames people have, especially if performance in certain domains is consistent across subjects, but violates logical principles. However, at present the frame view does not allow one to predict in advance which content areas will enhance performance. If one presumes that frames are built by domain general cognitive processes, then, at most, frame theory predicts that performance with the same content domain will vary, reflecting the idiosyncratic experiences of the subjects tested. But before frame theory can be considered a thoroughgoing explanation of content effects on the Wason selection task, the question of how frames are built must be addressed.

Auxiliary Mechanisms

In the early 1970s, several researchers considered the possibility that people are logically competent, but that abstract terms or relations create performance limitations (a "family 3" explanation). Wason & Shapiro (1971), Bracewell & Hidi (1974), and Gilhooly & Falconer (1974), suggested that thematic terms or relations may be more easily manipulated by auxiliary mechanisms (like working memory) than abstract terms or relations. The concrete terms used in thematic problems might enhance performance because they are more memorable than abstract symbols. A thematic relation might impose a smaller "cognitive load" on working memory if its content activates knowledge that cues the subject that the relation is non-reversible:* the fact that the conditional is not reversible need not be activated as a separate and additional piece of information in working memory.

Research into this view was virtually abandoned as later results called into question the very existence of a thematic content effect. The suggestion that the superior memorability of concrete terms explains the content effect can be ruled out. The food problem has never elicited a content effect, the post office problem does not when subjects are unfamiliar with the relation, the school and transportation problems usually do not produce content effects -- yet all use concrete terms.

The hypothesis that certain thematic relations reduce cognitive load is unlikely given the spotty replication record

* It is obvious that "All horses are animals" does not imply that "All animals are horses"; it is not so obvious that "All cards with an A on one side have a 3 on the other side" does not imply that "All cards with a 3 on one side have an A on the other side."

for the transportation problem. The transportation problem is one of the only thematic relations tested that clearly does not imply its converse. "Every time I go to Boston I travel by car" is a rather ordinary claim about a habitual way of getting to a particular destination, but "Every time I travel by car I go to Boston" sounds like the car has a mind of its own. The transportation results from the late 1970s and early 1980s have cast doubt upon the claim that a thematic relation enhances logical performance at all -- a fact that must be established before entertaining hypotheses regarding how it does this.

Before anyone realized how "elusive" content effect on the Wason selection task would prove to be, two sets of researchers -- Bracewell & Hidi (1974) and Gilhooly & Falconer (1974) -- tried to assess the relative contribution of concrete terms versus concrete relations to success on the transportation problem. Their results were contradictory.

Both sets of researchers investigated four types of problem:

Abstract Terms - Abstract Relation (AT-AR): "If there is a J on one side then there is a 2 on the other side"

Abstract Terms - Concrete Relation (AT-CR): "If I go to J then I travel by 2."

Concrete Terms - Abstract Relation (CT-AR): "If Manchester is on one side then car is on the other side."

Concrete Terms - Concrete Relation (CT-CR): "If I go to Manchester then I travel by car."*

As mentioned in Chapter 2, Bracewell & Hidi also tested two different linguistic formats: "Every time P, Q" and "Q every time P."

All told, Bracewell & Hidi had eight groups (two linguistic

* The CT-CR and AT-AR rules correspond to Wason & Shapiro's (1971) thematic and abstract rules. These are the groups described in Chapter 2.

formats for each of the above four groups), with 12 subjects per group. Their results are pictured in Table 3.1 below:

Table 3.1 Results of Bracewell & Hidi, 1974

	CT-CR	AT-CR	CT-AR	AT-AR	Totals
Every time P, Q:	9	4	1	1	15
Q every time P:	2	3	0	1	6
Totals:	11	7	1	2	

Number of subjects who answered 'P & not-Q'; n=12 per cell.

Bracewell & Hidi found a main effect for the linguistic format factor ("Every time..." does better), a main effect for the relation factor (the concrete relation does better), but no effect for the term factor. However, a further analysis of their data throws doubt on whether a relation factor exists at all. Although Bracewell & Hidi's data are consistent with the hypothesis that there is a relation factor, two alternative hypotheses are more strongly supported by their data:

1. There is no relation factor. Performance is enhanced only for the CT-CR problem, and only in the "Every time" format (see contrasts L1 below),
2. A concrete relation is sufficient to enhance performance, but only in an "Every time" format (see L2).

Bracewell & Hidi's hypothesis that both the relation and the linguistic format factor are important is represented by the set of contrasts L3.

$$L1 = \begin{matrix} +7 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 \end{matrix} \qquad L2 = \begin{matrix} +3 & +3 & -1 & -1 \\ -1 & -1 & -1 & -1 \end{matrix} \qquad L3 = \begin{matrix} +1 & +1 & 0 & 0 \\ 0 & 0 & -1 & -1 \end{matrix}$$

The sum of squares for L3, Bracewell & Hidi's hypothesis, accounts for 62% of the variance due to main effects and

interactions (i.e., of $SS_{total} - SS_{error}$; $F_{1,88} = 22.79$, $p<.001$, effect size $r = .45$). However, hypothesis L2 (a concrete relation helps only in an "Every time" format) accounts for 69% of the variance ($F_{1,88} = 25.35$, $p<.001$, $r=.47$) and hypothesis L1 (performance is enhanced only by a CT-CR problem in an "Every time" format) accounts for 80% of the variance ($F_{1,88} = 29.40$, $p<.001$, $r=.50$).

Thus, the hypothesis that the only cell showing an enhancement in logical performance is the CT-CR "Every time" cell -- the cell that exactly duplicates Wason & Shapiro's thematic group -- accounts for 18% more of the variance to be explained than Bracewell & Hidi's hypothesis that the relation factor exercises a separate effect, independent of linguistic format.

The efficacy of a concrete relation is further called into question by Gilhooly & Falconer (1974), whose results exactly contradict Bracewell & Hidi's. Gilhooly & Falconer investigated only one linguistic format ("Every time P,Q"), but their experiment is otherwise identical to Bracewell & Hidi's. The percentage correct for Gilhooly & Falconer's four groups is shown in Table 3.2. These figures reveal a significant main effect ($p<.05$) for the term factor (concrete terms do better), but no main effect for the relation factor, and no interactions. This directly contradicts the results of Bracewell & Hidi, who found a main effect for the relation factor, but no effect for the term factor (L3).

Indeed, the limited support that Bracewell & Hidi found for a relation factor may have been nothing more than a procedural artifact. Their unusual instruction that the conditional is "not

Table 3.2 Results for Gilhooly & Falconer, 1974

	CR	AR	
CT	11	10	21
AT	6	3	9
	17	13	

Number of subjects who answered 'P and not-Q'.
n=50 per cell.

reversible" may have simply focused subjects' attention on the relation factor (see Chapter 2).

In short, Bracewell & Hidi and Gilhooly & Falconer provide no clear evidence for the claim that a thematic relation enhances logical performance at all, thus ruling out the hypothesis that it does this by reducing "cognitive load".*

In light of the evidence presented in Chapter 1 indicating that people do not use the basic inferences of the propositional calculus, explanations in terms of performance factors do not appear very promising. The data reviewed in Chapter 2 cast a pall on such an enterprise. Any future "performance limitation" explanations will have to explain 1) why some familiar, concrete content can be pushed through "auxiliary mechanisms" better than other familiar concrete content, and 2) why the same familiar, concrete content is sometimes processed easily, and sometimes only with great difficulty.

* One could argue that because an AT-CR rule uses abstract terms, it cannot cue non-reversibility; that it is not so clear that "Every time I go to J I travel by 2" does not imply "Every time I travel by 2 I go to J". If this were so, then the relevant cells for testing the relation factor are CT-CR versus CT-AR. Again, the results would be contradictory: these cells differ for Bracewell & Hidi, but not for Gilhooly & Falconer.

Family 2 explanations

The only family of explanation that has no representative in the literature is "family 2": Humans have a logic module, but it is only activated in answering questions within the framework of a well established theory of what is true of a content domain. On this view, people may use induction to generate hypotheses in unfamiliar domains, but once they develop some inductive confidence about their hypotheses, they test them deductively.

This explanation cannot account for the content effects reviewed in Chapter 2. Assuming that familiarity is some measure of a person's understanding of a domain, the logic module should switch on for familiar domains. How, then, could this theory account for the fact that some familiar domains elicit content effects but others do not (e.g., DAP v. food), and the same familiar domain sometimes produces an effect, and sometimes not (e.g., transportation, school)?

Other formulations are possible, but I can think of none that can handle the results of Chapter 2. For example, perhaps familiarity with the elements and relations in a domain is not enough; perhaps the logic module is activated only when the domain is familiar and the subject also has personal beliefs regarding the veracity of the relation expressed by a rule.* This explanation can also be ruled out.

One implication of this view is that people should be especially adept at evaluating the validity of conclusions when

---

* Even if this were true, it could not explain the results of Chapter 2. For example, the most robust and replicable content effect was for social contract problems. Yet they have no truth value; they are rules to be followed.

they have personal beliefs regarding their truth value. Van Duyne (1976), was interested in whether people reason more logically with sentences that express necessary truths or contingent truths. He asked 22 subjects to generate conditionals that they thought were "always true" (necessity condition) and "sometimes true" (contingency condition). Each subject solved two selection tasks that had been created from rules he himself had generated (one necessary, one contingent).

If a logic module is activated in answering questions within a well-established theory of what is true, then 1) Van Duyne's paradigm should produce a substantial amount of falsification (at least over 50%), and 2) performance should be better for "always true" conditionals than for "sometimes true" conditionals.* Neither prediction is borne out by the data. Levels of falsification were low: only 6 of the 22 subjects (27%) falsified for the "necessary truth", and 8 out of 22 (36%) falsified for the "contingent truth." These percentages are not significantly different, and even if they were, the inequality runs counter to prediction. In fact, if one requires that subjects not only give the correct answer, but give it <u>for the correct reasons</u> (as assessed by verbal explanations), subjects displayed far more insight into conditionals that were "sometimes true" than ones that were "always true".

Even more damning to this explanation is the considerable body of literature on "belief bias" (reviewed by Pollard, 1982), which indicates that people do not reason more logically when

---

* Insofar as one's theory of what is true in a domain is better established for rules which are "always true" than for those that are "sometimes true."

they have personal beliefs regarding the truth value of the conclusion (see section 2.3). In such cases, subjects' performance appears to be guided, in part, by a desire or tendency to confirm their personal beliefs. When the content of a conclusion agrees with a personal belief, they judge the argument valid, and when it disagrees, they judge the argument invalid. Pollard & Evans, 1981, have demonstrated this on the selection task. Using a paradigm like Van Duyne's (1976), Pollard & Evans found that subjects were much more likely to choose the not-Q card for conditionals that they thought were "usually" or "always" false, than for conditionals that they thought were "usually" or "always" true.

These were conditionals that subjects' had generated themselves. Hence, they were familiar and subjects' had opinions regarding their veracity -- optimal conditions for the activation of a logic module, according to the reformulated family 2 explanation. If a logic module is activated under these conditions, we should see a substantial amount of falsification in this experiment.

Although Pollard & Evans report selection frequencies for individual cards rather than for combinations of cards, the percentage of subjects who answered 'P & not-Q' can be estimated from the percentage of Q card selections.* At most, 8.5% of subjects falsified for "true" conditionals and 21% falsified for "false" conditionals -- hardly auspicious performance for an activated logic module. This result is fatal to the "familiarity plus veracity" formulation of the family 2 explanation.

* because no one who chose Q could have answered 'P & not-Q'.

3.3 Summary of explanations

The theories that have been proposed in the literature represent four of the five "families" of explanation listed at the beginning of this chapter:

Frame theory is a family 1-a explanation: Humans have no logic module; instead they use rules of inference that are appropriate to the domain suggested by the problem. Current formulations presume that frames are built by domain general information processing mechanisms.

Auxiliary mechanisms is a family 3 explanation: Humans have a logic module, but auxiliary mechanisms for manipulating information create performance limitations.

Mental models and **memory-cueing** are family 4 explanations: Humans have no logic module, just the ability to recognize contradiction when they see it. **Mental models** theory falls into category 4-a, as it proposes that people actively construct models of the premises in search of ones that will refute a tentative conclusion. **Memory-cueing** falls into category 4-b, as it proposes that a counter-example can become available only if a person has actually experienced one -- people do not actively construct mental models in search of refutation.

Differential availability is a family 5 explanation: Humans have no logic module; rather, their performance is guided by non-inferential, general purpose heuristics.

None of these theories is satisfactory. Some are too unspecified to evaluate against empirical evidence (frames, mental models). Others are better specified (differential availability, memory-cueing, auxiliary mechanisms), but cannot

account for important pieces of evidence. To try to account for this contradictory data, some of the theories add codicils that are either theoretically unsound (differential availability), have consequences that are refuted by existing data (differential availability, memory-cueing/reasoning by analogy), or must be interpreted so loosely as to render the theory completely untestable (memory-cueing/reasoning by analogy).

None of the theories explain why social contract rules are the only thematic content to consistently elicit large content effects.

* * *

The many results cited in Chapters 1-3 demonstrate that people do not have the sort of logic module necessary for Popperian-style everyday learning. The Wason selection task is particularly interesting because it is a test of our ability to to test hypotheses deductively. Although some of the theories presented in Chapter 3 provide accounts of how people can test hypotheses in the absence of a logic module (mental models, memory-cueing), these theories require that the individual bring a vast store of world-knowledge to the task.

This brings us back to the central problem: How do people acquire this world-knowledge? Is this knowledge accurate? Induction is usually conceived as a process by which the world imprints existing relations on our minds -- that is, the kinds of hypotheses it can be expected to generate describe relations between existing properties or elements.

Given that there are an infinite number of ways of carving the world into properties, and therefore an infinite number of

relations between properties to serve as hypotheses, we must generate an enormous number of incorrect inductions. Yet results on the Wason selection task show that we are very bad at testing descriptive rules -- the very sort of hypotheses that induction provides. How, then, do we weed out all these incorrect inductions?

More puzzling: If the evolutionary purpose of human learning is to provide valid generalizations about the world, then surely the need to detect violations is greatest for descriptive rules. Why, then, are we so bad at detecting violations of descriptive rules, but so good at detecting violations of social contract rules? Social contract rules do not describe the way things are; they do not even describe the way existing people behave. They prescribe: They communicate the way some people want other people to behave. They are rules to be followed. One cannot assign a truth value to them. Why, then, do we appear to reason logically in response to social contract rules, but not in response to descriptive rules?

These are the questions that motivate the remaining chapters.

Chapter 4

Darwinian Algorithms

4.1 Another view of human rationality

If adherence to the canons of formal logic is the measure of human rationality, then humans are not very rational. But there is another, teleological view of rationality: An organism is behaving rationally when it is behaving purposefully -- when it is employing means that are likely to accomplish its goal. By this criterion, humans may indeed be rational beings. On this view the question of human rationality becomes: Are our reasoning processes appropriate to the problems they were designed to solve?

* * *

Unless you are a creationist, you probably believe that the human mind -- like the rest of the body and its functions -- is the product of evolution.

This insight was of little value to psychologists at the time of William James, or when John B. Watson was debating William McDougall, because evolutionary theory was in its infancy -- it was too hazy, too imprecise.

This situation has changed dramatically in the last 20 years. Now, the dynamics of natural selection can be mathematically modeled with great precision. This allows evolutionary biologists to determine what kinds of traits will be quickly selected out, and what kinds of traits are likely to become universal and species-typical.

Consequently, evolutionary theory now can be used as a heuristic guide for psychological theory. This heuristic rests

on the recognition that natural selection has produced psychological mechanisms as responses to various selection pressures in a species' "environment of evolutionary adaptedness" (Bowlby, 1969) -- the environment in which the species evolved. The more important the adaptive problem, the more intensely selection will have specialized and improved the performance of the mechanism for solving it.

Our species spent over 99% of its evolutionary history as Pleistocene hunter-gatherers. During that time, the dynamics of natural selection should have operated in the production of information processing mechanisms just as they did in the production of morphological and physiological mechanisms. The Pleistocene savannahs are the human environment of evolutionary adaptedness; our cognitive processes should be adapted to it, not necessarily to the 20th century industrialized world.

Recently, a number of cognitive scientists -- Chomsky, Fodor, Marr -- have argued that the best way to understand any mechanism, either mental or physical, is to first ask what its purpose is, what problem was it designed to solve (e.g., Chomsky, 1975; Fodor, 1983; Marr & Nishihara, 1978).

That is exactly what evolutionary theory allows you to do -- it allows you to pinpoint what kind of problems the human mind should be very good at solving. And although it cannot tell you the exact structure of the algorithms that solve these problems, it can suggest what design features they are likely to have. It allows you to develop a "computational theory" for that problem domain: a theory specifying what functional characteristics a mechanism capable of solving that problem must have (Marr &

Nishihara, 1978).

From the point of view of evolutionary theory, it is very unparsimonious to assume that the human mind is a general purpose computer with domain general, content-independent processes. There are domains of human activity for which the evolutionarily-appropriate information processing strategy is complex, and deviations from this strategy result in large fitness costs. For such domains, humans should have evolved "Darwinian algorithms" -- specialized learning mechanisms that organize experience into adaptively meaningful schemas or frames. When activated by appropriate problem content, these innately specified "frame-builders" should focus attention, organize perception and memory, and call up specialized procedural knowledge that will lead to domain-appropriate inferences, judgments, and choices. Like Chomsky's language acquisition device, these inference procedures allow you to "go beyond the information given" -- to reason adaptively even in the face of incomplete or degraded information (Bruner, 1973).

There are many domains of human activity that should have Darwinian algorithms associated with them. Aggressive threat, mate choice, sexual behavior, pair-bonding, parenting, parent-offspring conflict, friendship, kinship, resource accrual and distribution, disease avoidance, and predator avoidance are but a few. Social exchange is another. The dynamics of natural selection rigidly constrain the kinds of social exchange that can evolve, providing insight into the structure of the mechanisms that regulate it. This structure, and its consequences for performance on the Wason selection task, will be explored in

Chapters 5 and 6.

4.2 A brief primer on natural selection

We came into existence through the process of evolution, and the single ordering process in evolution is natural selection. Therefore, whatever systématic properties we have were produced by natural selection. If we wish to understand these properties, we need to understand the process that produced them.

An allele is a gene that occupies a particular location (locus) on a chromosome. Only one allele can occupy one locus on a particular chromosome. Let's say that in a given population of individuals, 50% of the individuals in the population have allele A at a particular locus, and 50% have allele B at that locus. A and B are alternative alleles -- a given individual has either A or B at a particular locus, but not both. Evolution is a change in the proportion of alternative alleles in a population. Hence, if the ratio of A:B alleles changes from 50:50 to 60:40, then evolution has occurred. Evolution is, therefore, a zero sum game: any increase in the proportion of A alleles in the population comes at the expense of alternative alleles, like B.

Evolution can occur in only two ways. The proportion of alternative alleles can change either via random processes (such as genetic drift) or via natural selection. Natural selection is the process whereby an allele increases its representation in the gene pool by virtue of the effect it has on the individual carrying it. If allele A has an effect on its carrier that, for any reason, causes an increase the proportion of A over B alleles in the population, then natural selection has occurred -- A has

been selected for, and B has been selected against. Technically, "fitness" refers to genes, not to individuals. Allele A is considered more "fit" than allele B if A codes for a trait that increases the proportion of A alleles over the proportion of B alleles in the population.

But genes are located in individuals. This means that there is one, and only one, way that an allele can increase its relative frequency -- by coding for traits that enhance the reproduction of the individual carrying that allele, or of its relatives, who may also bear that particular allele. An allele's fitness is a direct function of the relative number of offspring produced who have a copy of that allele.

Selection does not occur for the "good of the species", the "good of the group", or even for the "good of the individual" -- in fact, it is not even clear what these expressions mean. Genes that code for traits that enhance their own replication will spread through the population, even if this eventually causes the species to become extinct (for excellent discussions see Williams, 1966, and Dawkins, 1982).

In short: Genes -- through the traits they influence -- can influence the rate of their own replication, and hence their frequency in subsequent generations. If they code for traits that influence their own replication positively, then there is positive feedback and their representation in the population increases; if they code for traits that influence their own replication negatively, then there is negative feedback, and their representation in the population decreases. Natural selection is the process by which positive and negative feedback

of genes on themselves regulate the existence and frequencies of those genes over generations. Genes that code for traits that tend to maximize their rate of replication -- their "inclusive fitness" -- will tend to spread through the population, displacing alternative alleles each generation, until the traits those genes code for become fixed in the population. When this happens, the trait coded for is called a "species-typical" trait.

By analyzing the dynamics of gene flow through populations in terms of natural selection theory, one can determine what kinds of traits are most likely to become species-typical. In cognitive psychology, such species-typical traits have been called "cognitive competences" (Chomsky, 1975).

"Adaptation" means something precise: An adaptation is an aspect of the organism that has come into being because it had the effect of promoting the frequency of the genes that code for it. Natural selection theory concerns itself with adaptive function: Why does one trait come to dominate a population rather than another? How susceptible is the population to invasion by mutant genes coding for a different trait? How does one trait compare with another in terms of replicating the genes that code for it? In speaking of mental algorithms as adaptations, one is asking: How does one information processing strategy compare to another, as measured in terms of the replication of the genes that underlie each?

The traits of an organism are there either as a result of random processes (the stochastic dimension of evolution) or because they were selected for because they promote the spread of the genes that code for them. This second process is natural

selection, and it is major shaping force of any ordered relation in organisms.

What does this mean for social exchange? If social exchange was important in human evolution, then the mechanisms that regulate it will have been shaped by natural selection. This means they must promote the spread of the genes that code for them. This has profound consequences for what kinds of strategies for engaging in social exchange can evolve, and these will be discussed in depth in Chapter 5.

Natural selection theory is not behavior genetics

Natural selection theory is not behavior genetics. Behavior genetics attempts to ascribe differences in the behavior of individuals to differences in their genes. In contrast, natural selection theory's primary utility is in discovering what sort of traits are likely to become universal and species-typical.* Differences in behavior between individuals are presumed to be the facultative responses of species-typical traits to differences in those individuals' environments.

All traits have a "genetic basis", and can therefore be acted on by natural selection; This claim has nothing to do with issues of nature "versus" nurture

Phenotypic traits must have a "genetic basis" if natural selection is to act on them. Although few people have difficulty accepting this premise when the trait in question is eye color, leg number, or height, they balk when the trait in question is a

* Except in cases of frequency dependent selection, in which a population is expected to manifest a balanced polymorphism.

psychological mechanism or a behavioral tendency. Many social scientists believe this premise entails vast and unwarranted assumptions about a hotly debated empirical issue: To what extent can environmental manipulations alter the phenotypic expression of a trait? They believe that evolutionary thinking requires a "nativist" stand on the "nature/nurture" issue. It does not.

If properly understood, the premise that traits have a genetic basis entails quite minimal claims that should be acceptable to any cognitive psychologist. In fact, these claims should be acceptable to anyone who is realizes that it is impossible, in principle, for a tabula rasa to learn anything (e.g., Hume, 1977/1748; Kant, 1966/1781; Quine, 1969; Popper, 1972).

Genes are the blueprint for the development of an organism. More precisely, they are molecules of DNA which, by virtue of their physical structure, organize surrounding molecules into an enveloping organism -- an entity that can replicate the DNA that built it. The phenotype is the manifest organism, the collection of morphological, physiological, mental, and behavioral properties -- however construed -- that make up the organism. Genotypes specify phenotypes. That is, by virtue of its molecular structure, the genotype carries "instructions" for building an organism in a given environment. Thus there is a causal link between genotype and phenotype.

Natural selection can be thought of as the process whereby environments shape the characteristics of organisms by affecting the frequency of alternative alleles. There are both constancies and regular variations in any species' environment of evolutionary adaptedness that natural selection can be expected

to take advantage of. Thus, there are two causal links between environment and phenotype: 1) environments select genotypes, which specify phenotypes; and 2) the genotype's instructions are environment specific -- the phenotype built "assumes" certain environmental characteristics. Change those characteristics, and you will very likely change the phenotype.

For example, the arrowleaf plant's environment of evolutionary adaptedness included both watery and dry habitats.* When the arrowleaf plant sprouts in water, it develops wide leaves; when a genetically identical clone sprouts on dry land, it develops narrow leaves. An arrowleaf plant whose genes could not respond facultatively to these regular variations in its environment would be at a selective disadvantage -- its seeds could prosper in one habitat, but not in the other.

However, the fact that the leaf's width varies with wetness does not mean that it varies with every environmental dimension, nor does it mean that its shape is infinitely plastic. Leaf width does not vary, for example, with the amount of poetry read to the plant. Similarly, there is probably no environmental factor that would cause the arrowleaf's leaves to grow into the shape of the Starship Enterprise.

Certainly it would be correct to say that "leaf width" has a "genetic basis" in the arrowleaf. The plant's genetic blueprint specifies how wide or narrow the leaf will be in various environments. Furthermore, the arrowleaf's genetic blueprint prevents its leaves from assuming the shape of the Starship Enterprise, and prevents them from being affected by poetry

---

* From a 1976 lecture by E.O. Wilson on norms of reaction.

readings. Thus, it would also be correct to say that the traits "failure to assume the shape of the Starship Enterprise" and "failure to be affected by poetry readings" have a "genetic basis" in the arrowleaf plant. However, the claim "Leaf width in the arrowleaf has a genetic basis, therefore it cannot be affected by variations in the environment" is clearly false. Leaf width has a genetic basis, yet a simple environmental manipulation -- planting it in a wet or dry habitat -- changes the width of the arrowleaf's leaves.

All traits have a "genetic basis." No matter how far from "direct" genic influence a trait seems to be, it is still built from structures and processes that were built from structures and processes that were, ultimately, specified in the organism's genetic blueprint.

One of the things a genetic blueprint does, however, is specify how a trait will develop under different environmental conditions. Traits differ in terms of which environmental factors can affect their development, and in what ways their development can be affected. This, then, is why evolutionary theory takes no position on nature/nurture questions: Although evolutionary theory requires the assumption that the physical, mental, or behavioral traits it discusses have a genetic basis, it frames no hypotheses as to how these traits may be affected by environmental manipulations. That is a question for physiologists, developmental biologists, and psychologists: for scientists who study the structure of the particular physical and mental mechanisms an organism has. Evolutionary biologists do not study mechanisms, they study questions of adaptive function

-- questions like why genes that code for leaf width that varies with wetness would outcompete genes that do not.

If one's goal is to change a trait in some particular way, one must understand the mechanisms by which environmental factors influence the phenotype. Environmental factors can influence the phenotype either by acting directly on the genes (mutations, viruses) or by acting on mediating structures or processes built according to the genes' specifications -- the phenotype. The mediating structures can be physical organs and physiological processes, innately specified mental structures and processes, or "higher level" algorithms constructed by learning processes constructed by innately specified structures and processes. Environmental factors cannot act on a vacuum. If you want to understand how certain kinds of information will affect learning, you have to study the mind's information processing mechanisms, and natural selection theory can help in this endeavor: the structures and processes that environmental factors act on are, ultimately, built by the genes, and were shaped by natural selection. But natural selection theory cannot, by itself, tell you which manipulations will produce which effects.

4.3 Why should Darwinian algorithms be specialized and domain specific?

> Nature has kept us at a great distance from all her secrets, and has afforded us only the knowledge of a few superficial qualities of objects; while she conceals from us those powers and principles, on which the influence of these objects entirely depends. Our senses inform us of the colour, weight, and consistence of bread; but neither sense nor reason can ever inform us of those qualities, which fit it for the nourishment and support of a human body.
>
> -- David Hume (1977/1748, p. 21)

Genes coding for psychological mechanisms that maximize the inclusive fitness of their bearers will outcompete those that do not, and tend to become fixed in the population. The maximization of inclusive fitness is an evolutionary "end"; a psychological mechanism is a means by which that end is achieved. Can a psychological mechanism be domain general and content-independent, yet realize this evolutionary end?

Consider how Jesus explains the derivation of the Mosaic code to his disciples:

> Jesus said unto him, "Thou shalt love the Lord, thy God, with all thy heart, and with all thy soul, and with all thy mind. This is the first and great commandment. And the second is like it, Thou shalt love thy neighbor as thyself. On these two commandments hang all the law and the prophets."
>
> Matthew 22: 37-40 (emphasis added)

Jesus has given his disciples a domain general, content-independent decision rule to be used in guiding their behavior. But what does it mean in practice? Real life consists of concrete, specific, situations. How, from this rule, do I infer what counts as "loving my neighbor as myself" when, for example, my neighbor's ox falls into my pit? Should I recompense him, or him me? By how much? How should I behave when I find my neighbor sleeping with my spouse? Should I fast on holy days? Should I work on the Sabbath? What counts as fulfilling these commandments? How do I know when I have fulfilled them?

In what sense does all the law "hang" from these two commandments?

It doesn't. That is why the Talmud was written. The Talmud is a "domain specific" document: an interpretation of the commandments that tells you what actions count as "loving God"

and "loving your neighbor" in the concrete, specific situations you are likely to encounter in real life. The Talmud solves the "frame problem" (e.g., Boden, 1977; Fodor, 1983) posed by a "domain general" rule like Jesus's.

A domain general decision rule like "Do that which maximizes your inclusive fitness" cannot guide behavior in ways that actually do maximize fitness, because what counts as fit behavior differs from domain to domain. Therefore, like the Talmud, psychological mechanisms governing evolutionarily important domains of human activity must be domain specific.

The easiest way to see that Darwinian algorithms must be domain specific is to ask whether the opposite is possible: In theory, could one construct a domain general, content-independent decision rule, that, for any two courses of action, would evaluate which better serves the end of maximizing inclusive fitness?

Such a rule must include a criterion for assessing inclusive fitness: there must be some observable environmental variable against which courses of action from any domain of human activity can be measured. The simplest variable that correlates with inclusive fitness is number of grand-offspring produced by the end of one's life. Using this criterion, the decision rule can be rephrased more precisely as, "Choose the course of action that will result in more grand-offspring produced by the end of your life."

But how could one possibly evaluate alternative actions using this criterion? Consider a simple, but graphic example: Should I eat feces or fruit?

Clearly, I do not have two parallel lives to lead for purposes of comparison, identical except that in one I eat feces and in the other, fruit. Will trial and error work? Although I do not know it, if I eat the feces, there is a good chance I will contract a disease and die -- a large fitness cost. And if I eat the fruit and do not die, I still do not know if I can eat feces: for all I know, feces could be a rich food source that would greatly increase my fecundity. Could I learn from others? If I watch some people eat fruit, and others eat feces, and notice that more feces-eaters than fruit-eaters die by some later point in time, how do I know whether their death was caused by eating feces or by one of the many other things they did before their illness? And why would <u>they</u> choose to learn this way when I do not? -- my population of "guinea pigs" would be selecting themselves out. Furthermore, if, like most Pleistocene hunter-gatherers, I am living among my close kin, their death through experimentation is also a fitness cost to me (see Chapter 5).

Perhaps I could smell both: I'll eat what smells good and avoid what smells bad. But this method violates the assumption that the information processing system is domain general, and side-steps the "grand-offspring produced" criterion entirely. Nothing smells "intrinsically" bad or good; feces smell just fine to dung flies. Moreover, why would I infer that foul-smelling entities should not be eaten? Admitting smell or taste preferences is admitting domain specific innate knowledge. Admitting the inference that foul-smelling or foul-tasting entities should not be ingested is admitting a domain specific innate inference.

Without domain specific knowledge like this, how would I possibly learn to avoid feces and ingest fruit? Even if this were possible, an individual with appropriate domain specific knowledge would enjoy a selective advantage over one who relies on "trial and possibly fatal error" (Shepard, 1985). The tendency to rely on trial and error in this domain would be selected out; domain specific Darwinian algorithms governing food choice would be selected for, and become a species-typical trait.

There is also the problem of deciding which courses of action to evaluate. The possibilities for action are infinite, and the best a truly domain general mechanism could do is generate random possibilities to be run through the inclusive fitness decision rule. When a saber-toothed tiger bounds toward you, what should your response be? Should you file your toenails? Do a cartwheel? Sing a song? Is this the moment to run an uncountable number of randomly generated response possibilities through the decision rule? And again, how could you compute which possibility would result in more grandchildren? The alternative: Darwinian algorithms specialized for predator avoidance, that err on the side of false positives in predator detection, and, upon detecting a potential predator, constrain your responses to flight, fight, or hiding.

The domain general "grandchildren produced" criterion fails even in these simple situations. How, then, could it work in more complicated learning situations, for example, when an action that increases your inclusive fitness in one domain decreases it in another? Suppose your domain general learning mechanism somehow allowed you to figure out that sexual intercourse is a

necessary condition for producing offspring. Should you, then, have sex at every opportunity?

According to evolutionary theory, no. There are large fitness costs associated with, for example, incest.* Given a potential partner with a physique, personality, or resources that would normally excite you sexually: The information that he or she is close kin must inhibit your sexual impulses.

Again, if you engaged in incest, and lost the baby after a few months, how would you know what caused the miscarriage? Your life is a series of many events (perhaps including sex near the time of conception with non-kin as well as kin), any one of which is a potential cause. Why conclude that sex with one individual, who physically and psychologically resembles other members of his sex in many respects, caused you to lose your baby?

The need to avoid incest implies the ability to spontaneously and automatically acquire the category "kin versus non-kin" by merely observing the world -- even if it were possible to learn it by engaging in incest, the fitness costs would be too high. But the "number of grand-offspring produced" decision rule cannot be used to acquire evolutionarily crucial categories through mere observation: unless a categorization

---

* Each person has, on average, four lethal equivalent genes: having only one lethal equivalent does not adversely affect your health, but individuals die when they are homozygous for one of these genes. The probability that a random individual has one of the same lethal equivalents as you is very small; however, the probability that a full sib shares a given lethal equivalent with you is 50%. If you only had one lethal equivalent, and mated with your full sib, on average, half your children would die in utero or at a very young age. As each person has about four, the selective cost is even higher. The reduced sexual recombination that attends inbreeding also imposes selective costs in long-lived species, having to do with parasite load (Tooby, 1982).

scheme is used to guide behavior, it has no consequences on fitness.

Kin recognition requires Darwinian algorithms tuned to environmental cues that are correlated with kin but not with non-kin. These cues must be used in a particular way: either they must be used to match self to other, as in facial or olfactory phenotype matching, or they must categorize others directly, as when one imprints during a critical period on those with whom one was raised -- this can be a reliable environmental cue in species where the individuals with whom one is raised are normally one's closest kin (Shepher, 1983). There are an infinite number of dimensions that could be used to carve the environment into categories; there is no assurance that a general purpose information processing system would ever hit on those useful for creating the kin/non-kin categorization scheme, and the "grandchildren produced" criterion cannot guide such a system toward the appropriate dimensions.

Then there is the problem of generalization. Suppose you were somehow able to figure out that avoiding sex with kin had positive fitness consequences. How would you generalize this knowledge about the kin/non-kin categorization scheme to other domains of human activity? Would you, for example, avoid any interaction with kin? This would be a mistake; selectively avoiding sex with kin has positive fitness consequences, but selectively avoiding helping kin has negative fitness consequences (given a certain envelope of circumstances, Hamilton, 1964).

Thus, not only must the acquisition of the kin/non-kin

categorization scheme be guided by domain specific Darwinian algorithms, but its adaptive use for guiding behavior is also domain specific. In the sexual domain, kin must be avoided; in the helping domain, they must be helped; when one needs help, kin should be among the first to be asked (Hamilton, 1964); when one is contagiously ill, kin should be selectively avoided (Tooby, 1982). The procedural knowledge governing how one behaves toward kin differs markedly from domain to domain. Only Darwinian algorithms with procedural knowledge specific to each of these domains can assure that one responds to kin in evolutionarily appropriate ways. Simply put, there is no domain general criterion of fitness that could guide an equipotential learning process toward the correct set of fit responses.

Trial and error learning is inadequate, not only because it is slow and unreliable, but because there is no domain-independent variable for signaling error. In the sexual domain, error = sex with kin. In the helping domain, error = not helping kin given the appropriate envelope of conditions. In the disease domain, error = infecting kin.

Consequently, there are only two ways the human mind can be built. Either:

1. All innate psychological mechanisms are domain general, and therefore do not track fitness at all,

or

2. Some innate psychological mechanisms are domain specific Darwinian algorithms with procedural knowledge specialized for tracking fitness in the concrete situations hominids would have encountered as Pleistocene hunter-gatherers.

Clearly, the first alternative is no alternative at all. Advocates of this position would have to explain how genes coding for traits that impede their replication could possibly outcompete genes that code for traits that enhance their replication. In other words, they must explain how a complex of maladaptive traits was able to displace a complex of adaptive ones.

Darwinian algorithms solve the "frame problem"

Darwinian algorithms can be seen as frame-builders, as learning mechanisms that structure experience along adaptive dimensions in a given domain. Positing them solves the "frame problem" -- which is another name for the objections to domain general mechanisms that were raised in the above discussion.

Researchers in artificial intelligence have found that trial and error is a good procedure for learning only when an organism already has a well-specified model of what is likely to be true of a domain, a model that includes a definition of what counts as error. Programmers call this finding the "frame problem" (e.g., Boden, 1977; Fodor, 1983). To move an object, make the simplest induction, or solve a straightforward problem, the computer must already have a sophisticated model of the domain in question: what counts as an object or stimulus, what counts as a cause, how classes of entities and properties are related, how various actions change the situation. Unless the learning domain is severely circumscribed and the procedures highly specialized and content-dependent -- unless the programmer has given the computer what amounts to vast quantities of "innate knowledge" -- the computer can move nothing, learn nothing, solve nothing. The

frame problem is a concrete, empirical demonstration of the philosophical objections to the tabula rasa. It is also a cautionary tale for advocates of domain general, content-independent learning mechanisms.*

Unfortunately, the lesson has been lost on many. Although most cognitive psychologists realize that their theories must posit some innate cognitive architecture, a quick perusal of textbooks in the field will show that these still tend to be restricted to content-independent operating system characteristics: short term stores, domain general retrieval and storage processes, imagery buffers. Researchers who do insist on the necessity of positing content-dependent schemas or frames (e.g., Minsky, 1977; Schank & Abelson, 1977), seldom ask how these frames are built. They seem to presume that frames are the product of experience structured only by domain general learning mechanisms -- yet the building of frames must also be subject to the frame problem. Even Fodor (1983), a prominent exponent of the view that the mind's innate architecture includes specialized, content-dependent modules, restricts these to what he calls "input systems": perceptual or quasi-perceptual domains like vision, hearing, and language. He doubts the existence of modules governing "central" processes like reasoning and problem solving. Yet one wonders: Without domain specific inference processes, how can all this perceptual data be expected to guide

---

* Darwinian algorithms specify inference procedures, and can therefore be seen as constraining the theoretical set of all possible inferences to a few that are useful to the organism. However, they are not "constraints on learning" -- indeed, it is not clear that an organism could learn anything at all without such "constraints." See Appendix A: "The Frame Problem and So-called Constraints on Learning."

our behavior in adaptive directions?

Restricting the mind's innate architecture to perceptual systems, a content-independent operating system, a domain general concept learning mechanism, a content-independent hypothesis testing procedure, and a small ragbag of dimensions for construing similarity, might be OK if it did not matter what a person learned -- if, for example, learning that E is the most frequently used letter in the English language were as critical to one's inclusive fitness as learning that a saber-toothed tiger can eat you for lunch. But what a person learns does matter; not only what, but when, how reliably, and how quickly. And even more important is what a person <u>does</u> with that knowledge. The purpose of learning is, presumably, to guide behavior. Should I eat gravel? Should I engage in incest? Should I give others the only food I have for feeding my children? When my brother and my cousin are equally in need, should I satisfy those needs equally? Natural selection theory provides definite answers to questions like these, because the wrong decision can result in large fitness costs. How can an equipotential learning system that simply looks for relations in the world, provide information about the relative value, in inclusive fitness terms, of alternative courses of action? It cannot; it has no standard for assessing it.

Cognitive psychologists can persist in advocating such systems only because they are not asking what problems the mind was designed, by natural selection, to solve. The Darwinian view is that humans have innately specified mental algorithms that allow them to pursue goals that are or once were correlated with

their inclusive fitness. These innately specified mental algorithms cannot all be domain general. Behavior is a transaction between organism and environment; to be adaptive, specific behaviors must be elicited by evolutionarily appropriate environmental cues. Only specialized, domain specific Darwinian algorithms can insure that this will happen.

* * *

The proposal that Darwinian algorithms guide inference on reasoning tasks qualifies as a "family 1-b" explanation: Although we do not have a content-independent logic module, for evolutionarily important domains, we have extensive networks of "hypotheses" about what is true and what is "relevant", as well as rules of inference to guide reasoning within the domain. These hypotheses and rules are innate, or else the product of "experience" structured by domain specific innate algorithms. For these domains we do not face -- ontogenetically -- the problem of weeding out an infinite number of incorrect inductions, because this has already been accomplished phylogenetically, over 4 billion years of evolution. Hence, we spontaneously generate only a small subset of the class of all possible hypotheses, and those that we do consider are likely to be true -- or if not true, then adaptively useful. Their purpose in not merely to allow us to describe the world, but to pick up and process the information that is most salient for guiding our behavior in adaptive directions.

According to equipotential meta-theory, content is noise. According to Darwinian meta-theory, content is signal. If there are content-dependent Darwinian algorithms that guide reasoning

in evolutionarily important domains, then different content domains will "call up" different rules of reasoning. Choices on reasoning tasks should vary systematically with problem content, so long as problem content involves evolutionarily important domains. Non-important domains should show no systematic variation, because was no selection producing mechanisms sure to get such problems "right"; at best, reasoning about such areas should be weakly patterned by whatever domain general mechanisms do exist. To test this view, I chose a domain of human activity that should have been adaptively patterned by natural selection, and that appears to elicit consistent and robust content effects on a reasoning task: social exchange. Chapter 5 analyzes how the dynamics of natural selection apply to social exchange, and what this allows one to infer about the characteristics of the Darwinian algorithms that regulate social exchange in humans.

Chapter 5

Human Social Exchange

Evolutionary biology provides a heuristic for guiding psychological theory; the research and theory presented in this dissertation is meant to be an illustration of its potential. This heuristic rests on the recognition that natural selection has produced psychological mechanisms as responses to various selection pressures. The more important the adaptive problem, the more intensely selection will have specialized and improved the performance of these mechanisms. Some of these mechanisms evolved to meet the adaptive problem of social exchange. Successfully conducted social exchange was a critically important feature of hominid evolution. Natural selection permits the evolution of only certain strategies for engaging social exchange. By studying the nature of these strategies, one can deduce many properties that human algorithms regulating social exchange must have, as well as much about the associated capabilities such algorithms require to function properly. Using this framework, one can then make empirical predictions about human performance in areas that are the traditional concern of cognitive psychologists: attention, communication, reasoning, the organization of memory, and learning. One can also make specific predictions about human peformance on reasoning tests like the Wason selection task.

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

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

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

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

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

* * *

5.1 Natural selection and social exchange:

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

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

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

Traits can be thought of as the embodiment of <u>strategies</u> for the propogation of the genes that code for them. By analyzing the dynamics of gene flow through populations, one can determine what kinds of traits will quickly be selected out, and what kinds of traits are likely to become universal and species-typical. Formally, this analysis can be cast in terms of game theory: one

strategy is pitted against another in a race to see which one comes to dominate the gene pool. Such games can be mathematically modeled with great precision.* During the last 20 years, game-theoretic models of the dynamics of natural selection have proliferated in evolutionary biology. This process has led to a startling discovery: there are certain strategies that simply cannot be selected for (e.g., Hamilton, 1964; Williams, 1966; Maynard Smith, 1978; Dawkins, 1982). Furthermore, game-theoretic analyses can be used to specify what stategies are likely to be selected for, and what properties these strategies must have. This claim deserves an illustration from the literature of evolutionary biology.

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

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

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

Now, consider this excerpt from Hamilton, 1972:

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

"cooperation."

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

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

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

* Other models of social exchange are possible, but they will not change the basic conclusion of section 5.1: that reciprocation is necessary for the evolution of social exchange. For example, the Prisoner's Dilemma assumes that enforceable threats and enforceable contracts are impossibilities (Axelrod, 1984), assumptions that are frequently violated in nature. The introduction of these factors would not obviate reciprocation -- in fact, they would enforce it.

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

Figure 5.1 Payoff Schedule, Prisoner's Dilemma

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

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

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

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

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

Let's say you and I are playing a one move Prisoner's Dilemma game. I would reason thus: "You will either cooperate or defect. If you cooperate, then I get a higher payoff by defecting, because T, the Temptation to defect, is greater than R, the reward I would get for mutual cooperation. If you defect, then I get a higher payoff by also defecting, because P, the

Punishment for mutual defection, is greater than S, the Sucker's payoff I would get if I cooperate and you defect. Therefore, no matter what you do, I am better off defecting." Your reasoning process would be identical, so we would both defect, and we would both get P, the Punishment for mutual defection. Let's say the payoff matrix in Figure 5.1 represented dollars: if you cooperate, I get $5 for defecting instead of $3 for cooperating. If you defect, I lose nothing by defecting instead of losing $2 by cooperating.

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

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

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

Figure 5.2 Social exchange sets up a Prisoner's Dilemma

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

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

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

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

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

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

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

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

4. There is mounting evidence that a baboon male forms "special relationships" with a few lactating (and therefore infertile) females and their infants: he protects them from conspecifics and predators in exchange for sexual access when the females wean their infants and become fertile again (e.g., Smuts, 1982; Strum, 1985). His repayment, by necessity, comes at a much later time.

The opportunity for on-the-spot repayment is rare in nature for several reasons:

1. The "items" of exchange are frequently acts that, once done, cannot be undone (e.g., protection from attack, alerting others to the presence of a food source);

2. Opportunities for simultaneous mutual aid are rare because the needs and abilities of organisms are continually shifting: the female baboon is not fertile when her infant needs protection, yet this is when the male's ability to protect is of most value to her;

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

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

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

Selection pressures change radically when individuals play a series of Prisoner's Dilemma games. Mutual cooperation -- and therefore social exchange -- can emerge between two players when 1) there is a high probability that they will meet again, 2) neither knows for sure exactly how many times they will meet,**

---

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

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

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

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

---

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

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

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

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

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

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

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

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

CONTENDER'S SCORES:

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

the particular strategies it was pitted against in Table 5.1.

Robert Axelrod conducted a round robin computer tournament

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

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

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

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

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

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

The details of TIT FOR TAT are not what is important about this story. The key point, which TIT FOR TAT illustrates, is that a cooperative strategy can invade a population of non-cooperators if, and only if, it cooperates with other cooperators and cheats on cheaters. Indiscriminate cooperation cannot be selected for in any species. We humans have the ability to cooperate for mutual benefit. This capacity could not have evolved unless it included algorithms for detecting -- and being provoked by -- cheating.

5.2 Social exchange and the Pleistocene environment:

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

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

The intellectual capacities of early hominids allowed them to generate many situations for which cooperation paid off. The most important of these was the capacity to make and use tools, and the capacity to generate novel behavioral procedures to

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

achieve a goal. The exploitation of a new savannah and woodland niche -- made possible by tool use -- allowed individuals to acquire food items too large to be consumed by a single individual (Tooby & DeVore, 1985).* This created the perfect opportunity to provide a large benefit to another individual at a very low cost to oneself. There is virtually no cost to sharing food that you cannot consume anyway, and tomorrow you may be the one who has found no food. Fossil evidence indicates that Pleistocene hunter-gatherers, like their modern counterparts, engaged in extensive food-sharing (e.g., Issac, 1978). Similarly, the cost of sharing tools is low compared to the benefits one can garner through using them -- and the cost of sharing information about tool making may be even lower.

When combined with their capacity to opportunistically manipulate the environment through tool use, our ancestors' ability to generate novel behavioral procedures** created situations in which coordinated, cooperative behavior could produce vast payoffs. Perhaps one of the best examples are the "profits" to be made through cooperative hunting. Acting together, several armed men can bring down a wooly mammoth; acting alone, a single armed man cannot.

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

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

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

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

Reconstruction of the exact causal chain that led to the evolution of cooperation is still a matter of debate (cf. Kinzey, 1985). The most important point is that the Pleistocene hunter-gatherer environment in which we evolved provided many opportunities for individuals to benefit from mutual cooperation.

The peculiarities of hominid behavioral ecology place some species-specific constraints on a computational theory of social exchange in humans. Exchange in most primates is restricted to relatively few "items": food, sexual access, defense, grooming. The fewer the items for exchange, the more "item-specific" the algorithms regulating exchange can (and should) be: What counts

as "error" -- cheating or under-reciprocating -- can be more closely defined, increasing the accuracy of one's mental accounting system and the accuracy of reference (see section 5.3). In contrast, human algorithms for regulating social exchange should be able to handle a wide and ever-changing array of "items" for exchange: tools, information about tool-making, participation in opportunistically-created, coordinated behavioral routines. This suggests that our algorithms for regulating social exchange -- and the associated cognitive capacities they require to function properly -- will have some human-specific properties. These will be discussed in the next section.

5.3. A computational theory of social exchange

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

For humans, an evolved species, natural selection in a particular ecological situation defines and constitutes "valid constraints on the way the world is structured" for a particular adaptive information processing problem. In the case of social

exchange, the ecological and game-theoretic aspects of hominid social exchange discussed above provide the ingredients for the construction of just such a computational theory. A computational theory of social exchange must be powerful enough to (1) permit the realization of a "possible" social exchange strategy, that is, a strategy that can be selected for, and (2) exclude "impossible" strategies, that is, strategies that cannot be selected for.

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

However, the grammar of social contracts does not exhaust the set of information processing problems posed by social exchange. The ability to successfully participate in social exchange also requires a number of other, associated cognitive capacities, some of which are necessary in a wide range of other evolutionarily crucial social interactions, like mating, pair-bonding, parenting, and aggression. Before progressing to the grammar of social contracts and its implications for performance on the Wason selection task, five associated cognitive capacities

entailed by social exchange will be examined:

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

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

5.3.1 Human social exchange requires some fundamental cognitive capacities.

Proposition 1. One must be able to recognize many different individual humans.

> The basic idea is that an individual must not be able to get away with defecting without the other individuals being able to retaliate effectively. The response requires that the defecting individual not be lost in a sea of anonymous others. (Axelrod & Hamilton, 1981)

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

Indeed, humans do seem to have a highly developed ability to recognize large numbers of different individuals. Recognition rates are over 90% for familiar faces that have not been seen for up to 34 years (Bahrick, Bahrick & Wittlinger, 1975). Patients with a lesion in a specific part of the right hemisphere develop a selective deficit in their ability to recognize faces,

---

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

called prosopagnosia (Gardner, 1974). Carey & Diamond (1980) present and review an impressive array of evidence from a wide variety of sources suggesting that humans have innately specified face-encoding schemas. We are also good at identifying individual human gaits (Cutting, Proffitt, & Kozlowski, 1978; Kozlowski & Cutting, 1977).

Proposition 2. One must be able to remember some aspects of the histories of one's interactions with different individuals.

First, one must be able to recognize that a previous interactant in a social exchange is, in fact, a previous interactant, and not, for example, a stranger, a mate, or an offspring. Second, once an individual has been identified as a previous interactant, information regarding whether that individual has been a cooperator or a cheater must become accessible to the decision procedures. Third, one needs an "accounting system" for keeping track of who owes who what. As discussed in section 5.1, most Pleistocene social exchanges involved "reciprocal altruism" -- exchanges in which reciprocation was delayed, not simultaneous. In a simultaneous, face-to-face exchange, if you see that the other person has come prepared to defect, you simply withhold what that person wants.* There is no need to remember how much you owe or are owed, because there is no owing: each transaction is either a complete exchange or a complete defection. The potential for cheating is much higher, however, in exchanges in which reciprocation is

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

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

The extent of the history of interaction that must become available to the decision procedure that regulates whether you agree to participate in a particular social exchange (and whether any of these facts need be consciously recalled) will depend on the details of the particular decision procedure humans have evolved. TIT FOR TAT requires only that the last transaction with each interactant be recalled. But TIT FOR TAT operates in a highly constrained and uniform universe where all transactions are simultaneous, the same payoff matrix applies to each transaction, and the size of the payoffs for both players is equal within each transaction. In contrast, payoff matrices in the real world are always in flux, and part of that flux is caused by the negotiative skills of the individuals involved. Moreover, violence is possible in the real world: exchange situations with individuals who can reliably use violence to get their way do not necessarily fit the constraints of a Prisoner's Dilemma. Thus, an algorithm better adapted to conditions in the real world might assess many more factors regarding one's past history with an individual, such as (1) the number of transactions one has had with that individual in the past, (2) how he behaved in those transactions, (3) the size of payoffs to

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

A decision procedure that used such data, current behavioral cues,* and the payoff matrix for the current interaction to compute the conditional probability that one's partner will cooperate, might be better adapted to the complexities of exchange in nature.** If so, then the need to take such factors into account has implications regarding the organization of human memory. Information about one's history of interaction with a particular person ought to be "filed" with that person, and activated quickly and effortlessly when an opportunity for exchange with that person arises. When the payoff matrix of the current

---

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

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

interaction is such that you will lose a great deal if I cheat you, then more of our past exchange history should become accessible than for trivial exchanges. When you believe I have cheated you in a major way, there should be a flood of memories about your past history with me: you must decide whether it is worth your while to continue our relationship. In addition, this information will help you negotiate with me if you choose to continue our relationship: You can communicate how large a cost I have inflicted on you now and in the past (so I can make amends if I want to continue the relationship), tell me how close you came to ending our relationship (i.e., categorizing me as a permanent defector), convince me that I have become increasingly untrustworthy, threaten to ruin my reputation by telling others about my past transgressions, and so on.

The activation of past situations in which I have cheated you may, in turn, activate other* affective mechanisms that communicate cost/benefit information: they may cause you to cry, turn your back on me, scream at me, hit me. The extent and nature of the overt aspects of your affective reaction communicates to me your view of the extent of my wrong doing: whether you view it as serious enough to require restitution, how much is required and how soon, whether you intend to cut me off if I defect again. Emotion communication can be viewed as one way individuals communicate cost, benefits, and behavioral

---

* I say "other" because I see no principled way of drawing a dividing line between emotion and cognition. The flood of memories you experience when I betray you is as much a part of your "emotional reaction" as your turning red and punching me out (see Tooby, in press; Tooby & Cosmides, in preparation). intentions to others in negotiative situations (see Cosmides, 1983).

Proposition 3. One must be able to communicate one's values to others.

To engage in an exchange with you, I must know what you want. Although language is certainly a useful means for communicating what one values, non-linguistic organisms can also engage in social exchange -- however, the range of items they can exchange is necessarily more limited. For example, chimps recruit support from others in aggressive encounters, and frequently form long-term coalitional relationships. These coalitions are social exchanges in which the exchanged "item" is mutual aid in fights. A chimp under attack bares its teeth, emits a fear scream, looks at the individual from whom it wants support, and holds out its hand, palm up, toward that individual. If the attacked chimp receives the requested support, its demeanor changes radically: its hair stands on end, it emits aggressive barks, and it charges its opponent -- looking over its shoulder frequently to see if its supporter is still with it. If the chimp does not receive support, it continues cowering with hair flat and teeth bared, screaming and holding out its hand to solicit support.

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

> This happened once after Puist had supported Luit in chasing Nikkie [another chimp]. When Nikkie later displayed [aggressively] at Puist she turned to Luit and held out her

> hand to him in search of support. Luit, however, did nothing to protect her against Nikkie's attack. Immediately Puist turned on Luit, barking furiously, chased him across the enclosure and even hit him. (de Waal, 1982, p. 207)

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

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

The ambiguity of reference in the absence of a shared referential system is no mere philosophical puzzle (e.g., Quine, 1969; Gleitman & Wanner, 1982). For example, it is not clear that the infliction of pain, in the absence of a shared referential framework, could communicate what it is that the individual inflicting the pain wants the other individual to stop doing. The difficulty of communicating desires in the absence of a shared system of reference is illustrated by certain "communication gaps" that occur between two different, but

closely related, species of baboons: hamadryas baboons and savannah baboons.

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

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

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

1. Because they are universally shared, emotion signals can be recognized by anyone. By aiding "translation", such signals expand the range of possible interactants to individuals who speak a different language and individuals who cannot yet speak a language (small children).

2. Emotion signals can function like intersubjective rulers, permitting an observer to scale the values of the person emitting the signal: A very loud scream indicates a greater cost to the screamer than a moderately loud scream. Signals like screams, smiles, and trembles are "analog": The louder the scream, the wider the smile, the more noticeable the tremble -- the more strongly the person can be presumed to feel about the situation causing her to scream, smile or tremble. Words do not provide such convenient rulers, precisely because they are arbitrary and discrete symbols.

Verbal expressions indicating size of cost or benefit are more "digital": One might reasonably use "very much" to describe the degree of one's desire in both these sentences: "I want very much for my child's cancer to go into remission" and "I want that apple very much" -- yet in these two cases the degree of desire is vastly different.

3. Emotion signals allow the incidental communication of values to potential interactants. By observing your emotional reactions to various situations, even though they are not directed at me, I can learn what you value, and hence what sort of exchange you are likely to agree to (see Proposition 4). The verbal alternative is a process akin to writing to Santa Claus: Reciting, or publicly posting a long list stating one's preference hierarchy...with periodic updates!*

However, the very properties that make a natural language a poor medium for communicating intensity of affect make it an excellent system for indicating "items" of exchange. The variety of "items" that can be exchanged is severely limited in a species that uses only emotion signals. Primates appear to exchange fight for fight, fight for sex, sex for sex, food for food, fight, or sex, groom for groom, groom for fight, food, or sex...and not too much else. The use of language does not, of course, eliminate the problem of ambiguous reference. In the absence of a shared referential semantics, knowing what a word refers to is no less problematic than knowing what a gesture refers to.** But a natural language permits a potentially infinite number of arbitrary, discriminable symbols to be

---

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

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

attached to a potentially infinite number of discriminable classes or entities. As new situations arise, new words can be opportunistically created to refer to them. Consequently, language permits a range and specificity of reference impossible in the purely gestural systems of most primates.

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

---

* At least for Homo sapiens sapiens. Homo erectus' tool kit stayed identical over a wide range of different environments -- from Asia to Africa -- for over 1.5 million years (Pilbeam, personal communication). Of course, this observation applies only to tools that are recognizable as such in the fossil record. For example, a branch used as a ladder would not show up in the fossil record.

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

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

Because the variety of items exchanged by nonlinguistic primates is so limited, each item could, in theory, have a unique cost/benefit weighting associated with it that is shared by most other members of the species (e.g., ten grooms deserves one fight, a season of protection by a male deserves exclusive sexual access at the height of estrus, etc.). In other words, each item could have a preprogrammed, universally acknowledged, "exchange rate."

But there can be no preprogrammed, universally acknowledged, "exchange rate" for a constantly changing array of tools and coordinated behavioral routines. Language combined with emotion signaling affords a uniquely powerful communicative system for social exchange in a planning, tool using, and opportunistically cooperative, species. A wide variety of items can be precisely specified through language, and their relative value to an individual can be simultaneously communicated -- either incidentally* or intentionally -- via emotion signals. Indeed, there is rudimentary evidence suggesting that some aspects of the

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

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

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

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

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

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

-- the cry of a friend should recruit more sustained attention than the cry of a stranger.

Not only should attention be drawn to emotion signals, but one's learning mechanisms should be quick to pick up what the signal refers to -- what, exactly, the person emitting the signal is reacting to. This implies that our referential semantics (see footnote, p. 33) includes "hypotheses" about what kinds of events emotion signals are likely to refer to -- hypotheses about what other individuals are likely to value. Having such hypotheses is all the more important because many negative emotion signals refer to valued items that are not present or have not happened, vastly complicating the task of assigning a referent. When I am hungry, I moan because the thing I value -- food -- is <u>not</u> present. You must infer my desire for food from my moan, even though there is no spatio-temporally contiguous event in which the signal (my moan) and the referent (food) are both present.

Evolutionary theory provides a rich heuristic base for developing theories about what kinds of preference information is included in our referential semantics. Because humans are tool users, planners, and cooperators who can invent many alternative means for realizing a particular goal, many specific items of human preference will differ from culture to culture in ways that depend on that culture's technology, political structure, and history. This does not mean, however, that desires are random. Evolutionary theory is rife with hypotheses regarding what states of affairs the typical human is likely to prefer -- a few of these are listed in Box 5.1. In addition to being very incomplete, this list is extremely simple, in that it assumes

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

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

Motivation scripts should be powerful and sophisticated, for the ability to model other people's values is useful in a wide variety of evolutionarily important social contexts, from social exchange to aggressive threat to mate choice to parenting. They should prove to be strong organizational factors in the construction and reconstruction of memories. Details that are normally considered insignificant should be more easily recalled when activated motivation scripts allow them to be perceived as causally linked to biologically significant variables.*

---

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

Box 5.1 Typical Human Preferences

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

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

Veridical recall of stories that violate the assumptions about human preference instantiated in our motivation scripts should be difficult. Motivation scripts should guide the reconstruction of such stories during recall, distorting the original story in ways that make motivational sense. Implicit motivational assumptions are so pervasive in human communication, that motivation scripts will probably be an essential component of any artificial intelligence program that can usefully converse in a natural language.

An emotion signal should not only recruit attention and activate one's own motivation scripts, it should arouse one's curiosity. One would expect increased tendencies to observe the emotion-arousing event and ask questions about it. Crowds gather around fights, children follow fire trucks to the scene of a fire, onlookers bombard police with questions at the scene of a crime. Journalists make a profession of gathering information about the values and behavior of people who have a large impact on our lives. Motivation scripts may guide inferences about what exactly a given emotion signal refers to, but it can do this only if it is fed concrete information. The concrete information one acquires by witnessing an emotion-arousing event fills in parameter values in motivation scripts, determining which data structures and inference procedures are appropriate in decoding the reacting person's values.*

---

* There are, of course, other good reasons for being curious about biologically significant events -- e.g., you yourself might be confronted with the same situation at some point. However, when such events impact potential interactants they should be <u>especially</u> interesting -- Nightly News coverage of a fire at your neighbor's house verus a fire in Charlestown; a fist fight in the halls of William James versus a fist fight in Southie.

Acquiring information about the values of potential interactants is, in itself, valuable. Decoding the value systems of potential interactants is therefore likely to become a cooperative enterprise in itself. We even have a name for such exchanges of information and "analysis" -- gossip. Gossip is usually about situations that cause emotional reactions in potential interactants -- exactly the kind of situations that provide a window into someone's values. The more biologically significant the information, the "hotter" the gossip: Events involving sex, pregnancy, fights, windfalls, and death should be particularly "hot" topics, especially when they signal a change in someone's needs, values, or capacity to confer benefits. Hot gossip should be particularly interesting and easily remembered. Gossip about people who can have a large impact on one's well-being should be especially interesting; gossip about people one does not know should be especially boring.

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

It would be useless for information about the preferences of different individuals to be stored together in a semantic network, filed under "preferences" or "values." Like information

about an individual's history of reciprocation, a model of an individual's preferences and motivations should be filed under his or her "name." When the opportunity to acquire more preference information about an individual arises, the model appropriate to that individual must be easily retrieved, not just any person-model, or a model of average preference. "Averaging" the fact that one person prefers Z to W but another person prefers W to Z into one model of "average" preference does not enhance one's ability to engage in social exchange.* Learning

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

is useless unless it increases your ability to make offers that maximally benefit you given the limits imposed by what Smith or Jones are willing to accept. Offering W to Smith is more likely to induce him to give you Y than offering him Z; exactly the reverse is true of Jones. If you value Z more than W, you are better off making Smith an offer; if you value W more than Z, then strike a deal with Jones. The proper decision can be made only if person-specific preference information can be conveniently retrieved.

Proposition 5. Human algorithms regulating social exchange should not be too closely tied to particular items of exchange.

That tools, information about tool making, and participation in opportunistically-created, coordinated behavioral routines were important items for exchange has implications for the

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

structure of human cognitive algorithms regulating social exchange. The more limited the range of items exchanged, the more specific the algorithms regulating exchange can be. For example, the items exchanged in a cleaning fish symbiosis can be directly specified in the algorithms regulating the exchange. The host fish is specifically programmed to discriminate cleaner fishes from similar looking prey items, and, upon recognizing one, to refrain from eating it. The cleaner fish is specifically programmed to discriminate a host fish from other large, predatory fish, and, upon recognizing one, to approach and eat its ectoparasites (Trivers, 1971). Whereas the exchange algorithms of other organisms can be specific to the relatively few items they exchange, human algorithms regulating social exchange should be able to take a wide variety of input items, as long as these items are perceived as costs and benefits to the individuals involved in the exchange.

However, some items should be more readily perceived as costs and benefits -- those for which the perceiver can ascertain a clear causal link to biologically significant variables like offspring, kin, sex, food, safety, shelter, protection, aggressive formidability, and dominance. For example, a Mr. Michael Pastore of Dallas recently made the following comment in an interview with The Wall Street Journal:

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

Mr. Pastore perceives a clear causal link between his "plastic

sliver" and a biologically significant variable: the ability to attract sexual partners. His perception that a Platinum Card can attract sexual partners is based, in turn, on the perception that owning one is causally linked to a variable that is biologically significant to females in choosing male sexual partners -- the ability to accrue resources.* Knowing this, one should readily assume that Mr. Pastore perceives owning an American Express Platinum Card as a benefit, and that if he did not own one he would probably be willing to give up other items in order to acquire one. It is a suitable item for social exchange.

5.3.2 The grammar of social contracts

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

Just as a grammar of the English language is a set of rules for distinguishing well-formed sentences from ill-formed sentences, a grammar of social contracts is a set of rules for distinguishing well-formed social contracts from ill-formed social contracts. It includes the set of assumptions about the rules governing social exchange that must somehow be incarnated in the psychological mechanisms of both participants. Without these assumptions, much of what people say, mean, and intend to do in exchange situations could not be understood or anticipated, because all the necessary specifications are not spelled out

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

directly in speech. This grammar creates the "cohesion of discourse" (Wason & Johnson-Laird, 1972, p. 92), and the cohesion of behavior, in interactions involving uncoerced exchange. It constitutes the procedural knowledge that individuals must share in order to communicate their intentions to others in this particular kind of negotiative interaction (see Cosmides, 1983).

Unlike the exchange algorithms of cleaner fishes or even baboons, human algorithms for regulating social exchange should be item-independent: they should represent items of exchange as costs and benefits to the participants, and operate on those represent-ations (see 5.2; Proposition 5). The proposed grammar of social contracts is therefore expressed largely in cost/benefit terminology.

The items valued by our hominid ancestors were correlated with costs and benefits in their inclusive fitness; otherwise social exchange could not have evolved. The strategic exigencies of exchanging items that had real effects on the inclusive fitness of the exchangers selected for algorithms programmed with a particular set of cost/benefit relations (see 5.1). These relations can be expected to regulate how we think about social exchange, even if the items we value today are no longer correlated with our inclusive fitness. The grammar of social contracts specifies these cost/benefit relations.

* * *

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

To make the discussion concrete, let's fill in some values for P and Q in the offer "If P then Q". Let's say I offer you

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

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

What conditions must hold for you to accept my offer?

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

Q -- receiving $1m from me -- must be something that you

consider to be a benefit. An "item" -- an act, entity, or state of affairs -- is a BENEFIT TO YOU (B(you)) if, and only if, it increases your utility above your zero level baseline.* Let's say you value having a million dollars (Q) more than you value not having a million dollars (not-Q); with a million dollars you could feed the starving masses, sail a yacht to Tahiti, whatever. Then Q -- having a million dollars -- constitutes a benefit to you. You will not accept my offer unless, at the time of acceptance, you believe that Q constitutes a benefit to you. Using terms defined with respect to your values (rather than mine), we can rephrase my offer as: "If P then B(you)."

An item is a COST TO YOU (C(you)) if, and only if, it decreases your utility below your zero level baseline.** In my offer, P -- approving my thesis -- is the item that I have made my offer of B(you) contingent upon. Usually, P will be something that you would not do in the absence of an inducement; otherwise, I would be incurring a cost (giving up Q, the million dollars) by making the offer (if you were going to approve my thesis anyway it would be silly of me to offer you the million dollars).*** If P is <u>not</u> something you expected to do in the absence of my offer,

---

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

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

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

then, in your value system, not-P (not approving my thesis) is part of your zero level baseline, 0(you). This means that if not-P comes to pass, you will not have moved from your zero utility baseline -- you will be no worse off than if my offer had never been made. Let's say that my thesis is terrible and approving such a work would violate your ethical standards, cause you to risk losing your tenure, be the first step in the downfall of Western civilization...whatever. Then P -- approving my thesis -- decreases your utility and is therefore a cost to you, C(you).

Stated in terms of <u>your</u> value system, my offer can now be rephrased as "If C(you) then B(you)". But other conditions must hold before you will accept my offer. There is a constraint on the magnitudes (absolute values) of B and C, namely, B(you) > C(you), or, equivalently, B(you) minus C(you) > 0. We will call B(you) minus C(you) your "profit margin". For you to accept my offer, a million dollars must be more of a benefit to you than approving a terrible thesis is a cost. If this is not the case there would be no point in your entering into the contract; it would not increase in your utility. The greater the magnitude of B minus C (the greater your profit margin), the more attractive the contract will appear (an offer of one million dollars for approval of my thesis is more attractive than an offer of one thousand dollars). A contract that reversed this constraint (such that C>>B) sounds perverse. For example, I doubt anyone would be silly enough to make the offer: "If you break your arm then I'll give you a penny." In fact, Fillenbaum (1976) found that subjects consider such offers "extraordinary" 75% of the

time, compared to a 13% rate for offers that fit the constraints described above.

What conditions must hold for me to be willing to make an offer?

We can also consider the contract from the point of view of the person offering it, in this case, me. What conditions must hold for me to be willing to offer a contract? First, I must believe that not-P (your not approving my thesis) will come to pass if I do not make the offer. This means that not-P is a component of my zero level baseline: if not-P comes to pass, my utility level will not have changed. Second, I must want P -- in my value system, having my thesis approved must increase my utility, it must be a BENEFIT TO ME (B(me)). Third, not-Q -- not giving you $1m -- usually will be part of my zero level baseline, 0(me); if you do not accept my offer, I do not plan on giving you $1m, and if not-Q comes to pass, I will not have moved from my zero utility baseline.* Fourth, if not-Q is part of my zero baseline, then Q -- giving you $1m -- represents a decrease in my utility and is therefore a COST TO ME (C(me)). Fifth, like you, I will not enter into the contract (offer it in the first place) unless B(me) > C(me) (unless getting my thesis approved is worth more to me than relinquishing the million dollars).

In other words, I want P, and I am willing to give up Q to get you to do P; but I am not willing to give up Q without getting P. (I want you to approve my thesis and I am willing to

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

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

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

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

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

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

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

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

As you can see, P represents a different utility level to me than it does to you. Ditto for Q. In a well-formed social contract -- a contract that I am willing to offer and you are willing to accept -- the utility levels associated with P and Q are those shown in Table 5.2.

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

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

		Your point of view	My point of view
P	(you approve my thesis)	C(you)	B(me)
not-P	(you do not approve my thesis)	0(you)	0(me)
Q	(I give you my $1m)	B(you)	C(me)
not-Q	(I do not give you my $1m)	0(you)	0(me)

An offer is not entirely symmetrical, however. Suppose there were some way of equating value systems. Although B > C for both of us (or else we would not both agree to the contract),

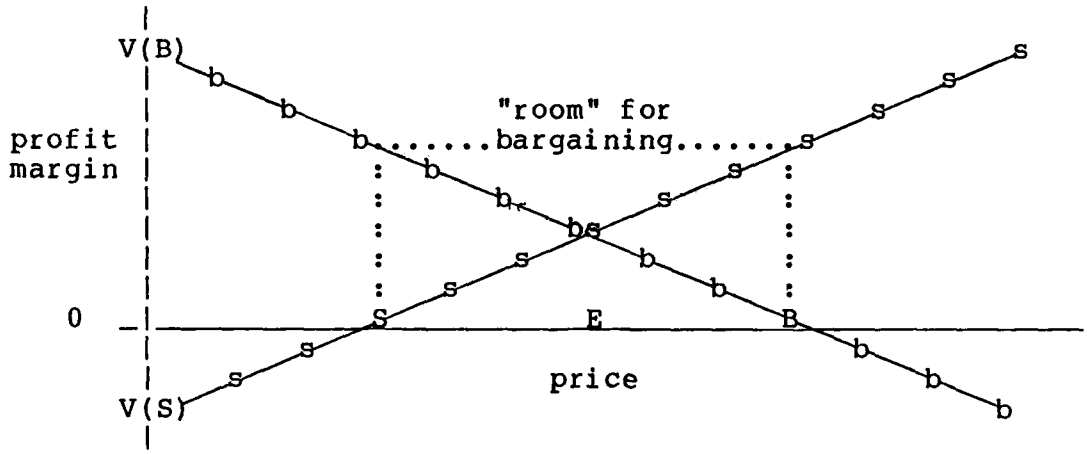
P (approving my thesis) might be a smaller cost to you than Q (giving up $1m to you) is to me (or vice versa). Likewise, Q might be a larger benefit to you than P is to me. These assymetries may lead to a difference in the magnitude of our profit margins (B minus C). Unequal profit margins invite bargaining: you attempt to increase your "profit margin" by paring down mine, and vice versa. Bargaining results in a zero sum game as long as both our profit margins are positive because more B(you) per unit C(you) corresponds to more C(me) per unit B(me). (See Figure 5.3; for a fuller account of these bargaining relations and their psychological sequelae, see Tooby, 1975). However, as long as your profit margin is greater than zero, it is in your interest to accept my offer, regardless of how large my profit margin is (and vice versa). If B > C for both of us, we have both benefited from the exchange. For this reason, I consider the term "subtle cheating", which Trivers (1971) uses to describe an interaction in which profit margins are unequal, to be a misnomer. "Under-reciprocating" is a more appropriate term; "cheating" should be reserved for the violation of a contract.

Snookering

> There is a joke that runs like this:
>
> A man from out of town walks up to a woman and says "If you sleep with me 3 times I'll give you $15,000." She is hard up for cash, so she agrees. After each session he pays her the money he promised. The woman decides this is an easy way to make money, so after she has been paid the full $15,000 she asks him if he would like to continue the arrangement. He says he can't because he must return home the next day. She asks "Where's home?" "Oshkosh," he replies. "Oh!" she says, "That's where my mother lives!" He answers, "Yes, I know. She gave me $15,000 to deliver to you."

The woman in the joke has been "snookered."

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



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

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

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

Man's offer: "If you sleep with me 3 times then I'll give you $15,000"

"If P then Q"

Woman's point of view: "If C(woman) then B(woman)"

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

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

"If P then Q"

Woman's point of view: "If C(woman) then 0(woman)"

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

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

Humor is frequently based on the violation of implicit assumptions. The punch line of this joke violates the woman's (and the listener's) implicit assumption that the man had offered a "sincere" contract. Above, we saw that when a contact is offered:

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

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

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

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

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

Likewise, the offerer has been snookered when:

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

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

Summary so far

The conditions that hold when an individual sincerely offers or sincerely accepts a social contract are shown in Table 5.3. For the sake of simplicity, P and Q stand for the actual items exchanged (these can be actions as well as entities). The first column shows the contract's cost/benefit structure in terms of the sincere offerer's value system; the second column shows what the sincere offerer believes the contract's structure is in terms of the acceptor's value system. The third column shows the contract's cost/benefit structure in terms of the sincere acceptor's value system; the fourth column shows what the sincere acceptor believes the contract's structure is in terms of the

Table 5.3

SINCERE SOCIAL CONTRACTS:
Cost/Benefit relations when one party is sincere, and that party believes the other party is also sincere.

	My offer: "If you give me P then I'll give you Q."			
	sincere offer		sincere acceptance	
	I believe:		You believe:	
P	B(me)	C(you)	B(me)	C(you)
not-P	0(me)	0(you)	0(me)	0(you)
Q	C(me)	B(you)	C(me)	B(you)
not-Q	0(me)	0(you)	0(me)	0(you)
profit margin	positive: B(me) > C(me)	positive: B(you) > C(you)	positive: B(me) > C(me)	positive: B(you) > C(you)
Translation:				
my terms...	"If B(me) then C(me)"		"If B(me) then C(me)"	
your terms...	"If C(you) then B(you)"		"If C(you) then B(you)"	

offerer's value system. The table shows that the sincere offerer and the sincere acceptor view the contract's cost/benefit structure in exactly the same way.

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

Table 5.4

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

	My offer: "If you give me P then I'll give you Q."			
	I try to snooker you; You accept sincerely		You try to snooker me, I offer sincerely	
	If you knew what I knew, we would both believe:		If I knew what you knew, we would both believe:	
P	B(me)	C(you)	0(you)	0(me)
not-P	0(me)	0(you)	?	C(me)
Q	0(me)	0(you)	B(you)	C(me)
not-Q	?	C(you)	0(you)	0(me)
profit margin	positive: B(me) > C(me)	negative: C(you)	positive: B(you) > C(you)	negative: C(me)
Translation:				
my terms...	"If B(me) then 0(me)"		"If 0(me) then C(me)"	
your terms...	"If C(you) then 0(you)"		"If 0(you) then B(you)"	

Social contracts as "Speech Acts"

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

Speech act theory is a part of analytic philosophy that grew out of the realization that, in speaking, people frequently do more than simply refer to something in the world. Frequently they do something by virtue of saying something. When I say "I promise to X", for example, I am not referring to something in the world: I am making a promise, and thereby incurring certain obligations -- I have committed a "speech act" (e.g., Searle, 1971). "Offering a contract" and "accepting a contract" can both be considered speech acts. Thus, we can ask the question, "What do I mean when I say 'If you give me P then I'll give you Q" and what do you mean when you say you "accept" my offer. Grice (1957,1967) has provided a convenient structure for understanding the meaning of speech acts.

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

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

1. I want you to give me P,

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

3. I realize, and I intend that you realize, that 4-9 are entailed if, and only if, you accept my offer:

4. If you give me P, then I will give you Q,

5. By virtue of my adhering to the conditions of this contract, my belief that you have given (or will give) me P will be the cause of my giving you Q,

6. If you do not give me P, I will not give you Q,

7. By virtue of my adhering to the conditions of this contract, my belief that you have not given (or will not give) me P will be the cause of my not giving you Q,

8. If you accept Q from me, then you are obligated to give me P (alternatively, If you accept Q from me then I am entitled to receive P from you),

9. If you give me P, then I am obligated to give you Q (alternatively, If you give me P then you are entitled to receive Q from me).

These rules capture the intercontingent nature of social exchange: they specify the ways in which the behavior of one person is contingent on the behavior of another person. Some philosophical niceties are discussed in Box 5.2 -- however, these points are not essential to the rest of the chapter.

Offering a contract is somewhat more complicated than other speech acts (like promises, see Searle, 1971) in that none of the conditions apply unless the hearer <u>accepts</u> the contract. In contrast, the conditions for a promise hold regardless of whether the hearer consents. Making a promise is a unilateral act; making a contract is not.

---

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

Box 5.2 Some Philosophical Niceties (categorized by clause)

2. In other words, the cost/benefit requirements do hold for me and I believe that they hold for you. (Note: sincere cost/benefit requirements entail "I value getting P from you more than I value keeping Q," so this need not be added as a separate statement.) Clause 2 is an implication of my offer even if the sincere cost/benefit requirements do not hold. After all, snookerers mean their offer to be thought sincere.

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

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

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

The correct notion of "cause" refers to the psychological realization of (the algorithm instantiating) this computational theory and the fact that it is guiding my behavior. My belief that you have given me P fills in the parameter value in the algorithm; this triggers the set of procedures within the algorithm corresponding to the contract's conditions of satisfaction. Triggering these procedures results in my giving you Q. This is the same sense of "cause" as in a computer program: the information that P can cause a computer to do something by virtue of that information's functional relation to various of its procedures. Let's say I have written a program in Basic instantiating all the conditions for making a social contract. The program then offers -- "If you type 'P' into me then I'll print 'Q' for you" -- and I accept. Part of the program would involve the computer waiting for me to fulfill my obligation, and this part may be written thus:

```
10 Input "Now give me P";A$
20 If A$ = P then go to 40
30 go to 10
40 Print "Q"
```

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

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

At first blush it might seem that a contract actually expresses a biconditional "Q if and only if P", and will therefore have the same truth table (see Chapter 1 for the truth table of a biconditional). If this were the case, the terms of the contract would be violated (someone would have cheated) if you are not in possession of Q after having done P (I cheated you), or if you are in possession of Q without having done P (you

cheated me). But it is not actually a biconditional because a social contract involves the twin notions of obligation and entitlement.

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

1. You have agreed to do P for me under certain contractual conditions (like 1-9), and

2. Those conditions have been met, and

3. By virtue of your not thereupon doing P, you agree that if I use some means of getting P (or its equivalent) from you that does not involve getting your voluntary consent, then I will suffer no reprisal from you.

Alternatively, 3 can be:

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

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

To take reprisal against someone trying to claim "just" restitution or punishment is to indicate that you are no longer interested in continuing a relationship with that person. In the contretemps between Puist and Luit, the two chimps discussed in

Proposition 3, Luit allowed Puist to punish him for his defection. I say "allowed" because Luit is far stronger than Puist, and in a pure test of strength Puist would not have a chance against Luit. To do otherwise would have signaled a drastic change in their several year reciprocal relationship.

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

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

As in obligation, an alternative formulation of 3 is:

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

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

A social contract is not a biconditional because I must do that which I am obligated to do, but I am not required to accept that to which I am entitled. If I pay the cost that I am obligated to pay (C(me), which corresponds to B(you)), I have fulfilled my end of the contract; I do not have to accept the benefit (B(me)) I am entitled to (however, you must offer it). Failure to accept a benefit one is entitled to may be foolish (and rare -- such behavior would have been strongly selected against), but it does not violate the terms of the contract.

Looking for cheaters

Cheating is the violation of the conditions of a social contract. It is the failure to pay a cost to which you have obligated yourself by accepting a benefit. The social contract can be explicit or implicit,* a private agreement or a law of your social group.

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

Let's say I offered, and you accepted, the following contract:

"If you give me P then I'll give you Q."

In your value system this translates to:

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

You have cheated me when you have accepted the item that corresponds to B(you) (item Q) without giving me the item that corresponds to C(you) (item P). In other words, you have cheated me when you have accepted item Q from me, but you have not given me item P. This means I have paid C(me) (item Q), but have not

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

received B(me) (item P). Your payoff: B(you). My payoff: C(me).

In my value system, the same contract translates to:

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

I have cheated you when I have accepted B(me) (item P) without paying C(me) (item Q). In other words, I have cheated you when I have accepted item P from you, but have not given you item Q. This means you have paid C(you) (item P), but have not received B(you) (item Q). Your payoff: C(you). My payoff: B(me). These relations are summarized in Table 5.5.

Table 5.5 How do you and I make out when one of us cheats the other?

	I cheat you		You cheat me		Contract fulfilled	
You give me P	B(me)	C(you)	---	---	B(me)	C(you)
You do not give me P	---	---	0(me)	0(you)	---	---
I give you Q	---	---	C(me)	B(you)	C(me)	B(you)
I do not give you Q	0(me)	0(you)	---	---	---	---
My payoff:	B(me)		C(me)		B(me)-C(me)	
Your payoff:	C(you)		B(you)		B(you)-C(you)	

As mentioned in Proposition 5, social contract algorithms in humans should be item-independent; they should represent items of exchange as costs and benefits to the participants, and operate on those representations. One cannot look out for cheating unless one can model the exchange's cost/benefit structure from the point of view of one's partner, as well as from one's own point of view.

This means that for any given exchange, two descriptions of

each item must be computed by the social contact algorithms. For a sincere contract, "If you give me P, then I'll give you Q", item P should be described as both B(me) and C(you), and item Q should be described as both C(me) and B(you) (see Table 5.5). The cost/benefit structure to oneself should be easily recoverable, even if the contract is phrased in terms of the value system of one's exchange partner.* There is an analogy here with the grammar of a language. The surface structure is the way the offer is actually phrased; the deep structure is a cost/benefit description of the surface structure from the point of view of each participant. The deep structure of the offer incorporates the information shown in Table 5.3 (or 5.4, if one person is snookering). One would expect these cost/benefit structures to be the descriptions from which participants construct paraphrases and reconstruct the course of the interaction from memory.

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

1. Ignore individual X if X has NOT accepted B(X)
2. Ignore individual X if X has paid C(X)
3. Watch out for individual X if X has accepted B(X)
4. Watch out for individual X if X has NOT paid C(X)

---

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

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

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

* * *

I doubt that most people would guess that the structure of a simple, straightforward social exchange is as complex as this chapter shows it to be. But then, that is a prediction of the theory. People usually do not realize how complex the grammar of their language is, yet they produce grammatical sentences with ease. Similarly, people do not realize how complex engaging in social exchange is, yet they do it with ease. Both parties implicitly understand and act on all the relations involved because both possess the same Darwinian algorithms for reasoning about social exchange.

*

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

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

Box 5.3 Exchange of "tools" in the Computer Age

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

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

And on "cheaters"...

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

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

Chapter 6

Social Contracts and the Wason Selection Task: Experiments

The game-theoretic conditions governing reciprocation and non-reciprocation in social exchange are too complex and too important to leave to the vagaries of trial-and-error learning. Humans should have evolved inferential procedures that make them very good at detecting cheating on social contracts. The "look for cheaters" procedure predicted in Chapter 5 can be expected to generate a quite specific and unusual pattern of responses on the Wason selection task when its content involves social exchange.

In a social exchange situation for which a subject has incomplete information, a "look for cheaters" procedure should draw attention to any person who has NOT paid the required cost (has he illicitly absconded with the benefit?) and any person who has accepted the benefit (has he paid the required cost?).

The Wason selection task is a paper and pencil problem that invites a subject to see if a conditional rule of the form "If P then Q" has been violated by any one of four instances (represented by cards) about which the subject has incomplete information (see Chapter 2). By presenting one term of a conditional rule as a rationed benefit, and the other term as a cost/requirement, one can create a Wason selection task that instantiates a social contract. This can be used to see how people reason about social contracts in the face of incomplete information. Figure 6.1 shows the cost/benefit structure of such a Wason selection task.

Figure 6.1

Structure of Social Contract (SC) Problems

It is your job to enforce the following law:

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

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

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

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

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

Irrespective of logical category, a "look for cheaters" procedure would cause the subject to:

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

As Figure 6.1 shows, the logical category to which each card corresponds varies, and is determined by where the costs and benefits to the potential cheater are located in the "If-then" structure of the rule. For a "standard" social contract (STD-SC) -- one with the benefit in the "If" clause and the cost in the "then" clause -- the two chosen cards correspond to the logical categories 'not-Q' (cost NOT paid) and 'P' (benefit accepted). However, for a "switched" social contract (SWC-SC) -- one with the cost in the "If" clause and the benefit in the "then" clause -- the same two cards correspond to the logical categories 'not-P'

(cost NOT paid) and 'Q' (benefit accepted). In Figure 6.1, Rule 1 is a STD-SC and Rule 2 is a SWC-SC.

The correct formal logic response is 'P & not-Q', regardless of content. Therefore, a subject using a "look for cheaters" procedure would appear to be reasoning <u>logically</u> -- by choosing 'P & not-Q' -- on STD-SC rules, yet appear to be reasoning <u>illogically</u> -- by choosing 'not-P & Q' -- on SWC-SC rules. In the literature, a rule is said to have produced a content effect when it elicits a higher percentage of logically falsifying, 'P & not-Q' responses than an abstract problem. By this definition, a STD-SC should elicit a content effect but a SWC-SC should not.

A "look for cheaters" procedure should ignore the "cost paid" card and "benefit NOT accepted" card. These correspond to 'P & not-Q' (the correct formal logic response) for a SWC-SC rule, and to 'not-P & Q' for a STD-SC rule. Thus social contract theory predicts that the dominant response to a STD-SC problem will be <u>very</u> rare on a SWC-SC problem, and vice versa.

To sum up, the correct SC answers to STD-SC and SWC-SC rules differ from the logically correct answers:

	SC answers		Logical answers	
	'P & not-Q'	'not-P & Q'	'P & not-Q'	'not-P & Q'
STD-SC:	yes	no	yes	no
SWC-SC:	no	yes	yes	no

Therefore, by comparing performance on STD-SC and SWC-SC rules one can tell if reasoning is governed by a logical procedure or a "look for cheaters" procedure.

If people do, in fact, have Darwinian algorithms governing

how they reason about social exchange, these ought to function, in part, as frame-builders that structure new experiences. This means they should operate in unfamiliar situations. No matter how unfamiliar the relation or terms of a rule, if the subject perceives the terms as representing a rationed benefit and a cost requirement -- that is, if the subject recognizes the situation as one of social exchange -- a "look for cheaters" procedure should produce the above pattern of responses. Non-social contract rules, either descriptive or prescriptive, should not show this particular pattern of variation, regardless of their familiarity. In general, they can be expected to elicit the same low levels of 'P & not-Q' and very low levels of 'not-P & Q' typically found in the literature for non-SC problems.

Previous results on the Wason selection task are consistent with a social contract interpretation (see Chapter 2 for a detailed review). Robust and replicable content effects are found only for rules that relate terms that are recognizable as benefits and costs in the format of a standard social contract. No thematic rule that is not a social contract has ever produced a content effect that is both robust and replicable. For thematic content areas that do not express social contracts, either no content effect is found (e.g., the food problem), or there are at least as many studies that do not find content effects as there are studies that do (transportation and school problems). Moreover, most of the content effects reported for non-SC rules are either weak (Gilhooly & Falconer, 1974; Pollard, 1981), clouded by procedural difficulties (Bracewell & Hidi, 1974; Van Duyne, 1974), or have some earmarks of a social

contract problem (Van Duyne, 1974). All told, for non-SC thematic problems, three experiments have produced a substantial content effect (transportation: Wason & Shapiro, 1971; Bracewell & Hidi, 1974; school: Van Duyne, 1974), two have produced a weak content effect (transportation: Gilhooly & Falconer, 1974; Pollard, 1981), and 14 have produced no content effect at all (transportation: Bracewell & Hidi, 1974; Manktelow & Evans, 1979; Yachanin & Tweney, 1982; Griggs & Cox, 1982. food: Manktelow & Evans, 1979 (4 experiments); Brown et al., 1982; Reich & Ruth, 1982; Yachanin & Tweney, 1982; school: Yachanin & Tweney, 1982. non-SC post office: Golding, 1980; Griggs & Cox, 1982). The few effects that were found did not replicate. In contrast, sixteen out of sixteen experiments with standard social contracts elicited substantial content effects. These include the Drinking Age Problem, the Post Office Problem, and the Sears Problem. In this extensive literature, STD-SC rules are the only thematic content rules to elicit strong, replicable content effects on the Wason selection task.

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

The alternative hypothesis: Availability

"Availability" theory comes in a variety of forms with some important theoretical differences, but common to all is the notion that the subject's actual past experiences create associational links between terms mentioned in the selection task (see Chapter 3 for detailed explanations). The more exposures a subject has had to, for example, the co-occurrence of P and Q, the stronger that association will be, and the easier it will come to mind -- become "available" as a response. A subject is more likely to have actually experienced the co-occurrence of 'P & not-Q' for a familiar rule, therefore familiar rules are more likely to elicit logically falsifying responses than unfamiliar rules. However, if all the terms in a task are unfamiliar, the only associational link available will be that created between P and Q by the conditional rule itself, because no previous link will exist among any of the terms. Thus 'P and Q' will be the most common response for unfamiliar rules. Falsifying responses will be rare for all unfamiliar rules, whether they are social contracts or not.

Testing social contract theory against availability theory

At present, it is widely believed that some variant of availability theory accounts for all content effects on the Wason selection task. In contrast, social contract theory proposes that for content involving social exchange, a social contract algorithm is the primary regulator of responses. Thus, for social contract theory, the major determinant of responses is

whether a rule is a social contract (SC) or descriptive (D).* For availability theory, the major determinant of responses is whether a rule is familiar (F) or unfamiliar (U). Because these two variables are orthogonal, one can create an array of four problem types: Unfamiliar-Social Contract (U-SC), Familiar-Social Contract (F-SC), Unfamiliar-Descriptive (U-D), Familiar-Descriptive (F-D):

	'Availability dimension'	
	familiar	unfamiliar
descriptive	F-D	U-D
'SC dimension' social contract	F-SC	U-SC (U-STD-SC, U-SWC-SC)

Moreover, there are two kinds of U-SC problems: unfamiliar standard social contracts (U-STD-SC) and unfamiliar switched social contracts (U-SWC-SC). All but the F-SC, which confounds familiarity with being a social contract, can be used to construct critical tests disentangling the following two hypotheses:

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

* Actually, the non-SC rule can be either descriptive or prescriptive. All SC rules are prescriptive, but not all prescriptive rules are SC rules. Most of the non-SC thematic problems tested in the literature were descriptive.

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

It would be difficult to believe that availability has no effect on familiar problems. The SC hypothesis is silent on this point. Indeed, any effect availability might have in eliciting falsifying responses to familiar descriptive problems can be used as a metric for judging the size of a social contract effect. SC algorithms can be said to be a major determinant of responses for problems involving social exchange if there are more SC responses (STD-SC: 'P & not-Q'; SWC-SC: 'not-P & Q') to unfamiliar social contract problems than falsifying responses to familiar descriptive problems.

In the experiments that follow, story context was used to transform an unfamiliar (U) rule -- like, "If a man eats cassava root, then he must have a tattoo on his face" -- into either a U-SC or U-D problem. By embedding the same unfamiliar rule in two different stories, one can contextually define that rule as either a social contract (U-SC) or a descriptive rule (U-D). A social contract story contextually defines one term of the unfamiliar rule as a rationed benefit that must be earned and the other term as a cost/requirement. A descriptive story does not define the terms as costs and benefits, but it does contextually link them to familiar concepts and tie them together by a familiar relation. In these experiments, the U-D problems invoke what should be one of the most familiar relations according to standard associationism: spatio-temporal co-occurrence.

Switching the position that a U-SC's terms occupy in the

"If-then" structure of the rule transforms its theoretical status from standard (U-STD-SC) to switched (U-SWC-SC), or vice versa. Let's say a story portrays cassava root as a rationed benefit and having a tattoo as a cost requirement. Then:

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

has a standard SC format, whereas:

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

has a switched SC format. Switching the position of a U-D's terms does not change its theoretical status.

Experiments 1 through 4 are the most important from a theoretical point of view. They compare performance on unfamiliar social contract problems to performance on unfamiliar and familiar descriptive problems. These experiments permit six critical tests -- comparisons for which hypotheses **AV** and **SC** make radically different predictions. These tests address the following questions:

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

Experiment 5 tests whether an abstract problem with a social contract story context can elicit SC responses; Experiment 6 shows that the social contract effect is replicable with familiar content.

Experiment 1

The purpose of Experiment 1 was to see whether an unfamiliar STD-SC problem would elicit the predicted SC response, 'P & not-Q'. A high percentage of "falsifying" responses on a U-STD-SC is predicted only by social contract theory; availability theory predicts a low percentage of falsifying responses on the U-STD-SC because it is unfamiliar. Each subject was asked to solve four Wason selection tasks, which were presented in random order. Theoretically, the problem types can be described as follows:

U-STD-SC:	Unfamiliar - Standard Social Contract
U-D:	Unfamiliar - Descriptive
AP:	Abstract Problem
F-D:	Familiar - Descriptive

The AP was a non-SC prescriptive rule; it was included because it is commonly used as a standard for assessing availability (Wason, 1983).

Table 6.1 shows the relative percentages of 'P & not-Q' and 'not-P & Q' responses expected in Experiment 1, assuming that responses are determined by either SC algorithms or availability, but not both.

Subjects.

Twenty-four undergraduates from Harvard University participated in Experiment 1; they were paid volunteers recruited by advertisement (13 females, 11 males; mean age: 19.4 years).

Table 6.1 Predictions, Experiment 1: Social contract theory versus availability theory.

STANDARD Social Contract (STD-SC) v. Descriptive (D) problems.

	'P & not-Q'		'not-P & Q'
	Social contract	Availability	Both theories
U-STD-SC:	high	low	very low
U-D:	low	low	very low
AP:	low	low	very low
F-D:	low	middling to low	very low

Table 6.1 Relative percentages of 'P & not-Q' and 'not-P & Q' responses expected for Exp 1, assuming that responses are solely determined by either SC algorithms or availability.

'P & not-Q' responses: Social Contract Predictions: Rationale for U-STD-SC described in text. Because SC algorithms will not be structuring responses to non-SC problems (U-D, AP, F-D), these should not elicit high levels of SC responses. They can be expected to elicit the same low levels of 'P & not-Q' responses (and very low levels of 'not-P & Q', see below) typically found in the literature for non-SC problems. SC is silent on whether availability will exert an independent effect on F-D problems. Availability Predictions: The F-D transportation problem could elicit a somewhat higher percentage than the unfamiliar problems (U-STD-SC, U-D, and AP) because it is likely that at least some subjects living in the Boston area have a pre-existing or dominant 'P & not-Q' association (see Chapter 3).

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

Materials and Procedures.

Each subject received a sealed booklet with instructions on the first page, followed by four Wason selection tasks, one per page. Each selection task was embedded in a brief story. Each booklet contained a U-STD-SC, a U-D, an AP, and an F-D. The order of the four problems was randomized across subjects. Experiment 1 had a within subjects design.

All stories were phrased so as to activate a "detective set" (Van Duyne, 1974), and all asked subjects to look for violations of the rule. There were two versions ('A' and 'B') of unfamiliar problems; the rules used and the two cultures described therein were fictitious. A booklet either contained the 'A' version of the U-D problem and the 'B' version of the U-STD-SC, or vice versa. Figure 6.2 shows the 'A' versions of the unfamiliar problems; Figure 6.3 shows the 'B' versions. The 'A' version of the U-D and U-STD-SC varied only in surrounding story context; the rules used were identical. The same was true of the 'B' version problems. Figure 6.4 compares the unfamiliar problems used in Experiments 1 and 2.

I wanted to use any effect availability might have in eliciting falsifying responses to the F-D problem as a metric for judging the size of a social contract effect. For this reason I used a transportation problem as the F-D problem; the transportation problem had been the most successful non-SC problem in the literature (see Chapter 2). Various versions (using different terms) of the F-D and AP problems were randomized with respect to each other and the unfamiliar problems. Figure 6.5 shows examples of the F-D and AP problems used.

Figure 6.2 'A' versions of Unfamiliar rules

Unfamiliar Standard Social Contract (U-STD-SC)

PAGE

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

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

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

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

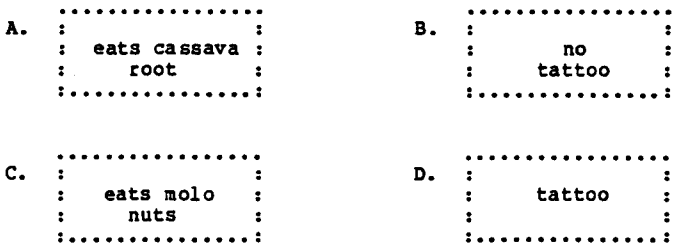
Although everyone craves cassava root, eating it is a privilege that your people closely ration. You are very sensual people, even without the aphrodisiacal properties of cassava root, but you have very strict sexual mores. The elders strongly disapprove of sexual relations between unmarried people,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

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

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

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

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



Unfamiliar Descriptive (U-D)

PAGE

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

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

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

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

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

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

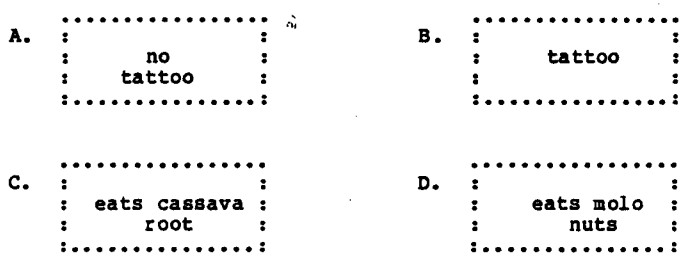


Figure 6.3 'B' versions of Unfamiliar rules

Unfamiliar Standard Social Contract (U-STD-SC)

PAGE

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

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

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

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

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

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

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

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

A.	eats some duiker meat	B.	has never found an ostrich eggshell
C.	does not eat any duiker meat	D.	has found an ostrich eggshell

Unfamiliar Descriptive (U-D)

PAGE

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

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

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

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

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

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

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

A.	quail eggshell	B.	ostrich eggshell
C.	duiker	D.	weasel

Figure 6.4 Experiments 1 and 2: Comparison of Unfamiliar Problems

'A' VERSIONS OF UNFAMILIAR PROBLEMS

Rule A:

(Exp 1): "If a man eats cassava root, then he must have a tattoo on his face."

(Exp 2): "If a man has a tattoo on his face, then he eats cassava root."*

Cards: (Exp 1 & 2)	Logical Category: (Exp 1)	(Exp 2)
"eats cassava root"	P	Q
"eats molo nuts"	not-P	not-Q
"tattoo"	Q	P
"no tattoo"	not-Q	not-P

A:Common Story Elements: All 'A' problems involve a fictional Polynesian people called the "Kaluame". A facial tattoo means a man is married. no tattoo means he is a bachelor.

A:U-D Story Summary: You (the subject) are an anthropologist studying the Kaluame. Cassava root and molo nuts are foods found on two different parts of their island; men eat one or the other, but not both. Married men and bachelors tend to live on two different parts of the island, where cassava root and molo trees grow, respectively. You read a report asserting that Rule A seems to hold. If so, perhaps men are simply eating foods which are most available to them. But Rule A may not be true. You want to investigate for yourself by seeing if any of four men (each card represents one man) are breaking Rule A.

A:U-STD, U-SWC Story Summary: You are a Kaluame who has been entrusted to enforce your people's laws. Cassava root is a powerful but scarce aphrodisiac and a delicious food source; eating it is a rationed privilege, governed by Rule A. Molo nuts are a common, undesirable food source with no interesting "medicinal" properties. The Kaluame have very strict sexual mores, and disapprove of sexual relations between unmarried people. Your job is to catch any of four men (each card represents one man) who might have cheated on Rule A.

*"must" is left out of the first clause because it violates common English usage36.

'B' VERSIONS OF UNFAMILIAR PROBLEMS

Rule B:

(Exp 1): "If you eat duiker meat, then you have found an ostrich eggshell."

(Exp 2): "If you have found an ostrich eggshell, then you eat duiker meat."

U-D Cards: (Exp 1 & 2)	U-STD, U-SWC Cards: (Exp 1 & 2)	Logical Category: (Exp 1)	(Exp 2)
"duiker"	"eats some duiker meat"	P	Q
"weasel"	"does not eat any duiker meat"	not-P	not-Q
"ostrich eggshell"	"has found an ostrich eggshell"	Q	P
"quail eggshell"	"has never found an ostrich eggshell"	not-Q	not-P

B:Common Story Elements: In all 'B' problems, you (the subject) are an anthropologist studying a (fictional) southwest African hunter-gatherer group called the "Namka". Duikers are antelopes whose meat the Namka eat.

B:U-D Story Summary: You want to know if Rule B is fact or folklore. The whereabouts of both duikers (hunted for their meat) and ostrich eggshells (used as canteens) are of interest to the Namka. Duikers frequently feed on ostrich eggs, so you guess Rule B reflects a real relationship that the Namka use to guide their foraging behavior. You don't know the Namka's language well enough to ask, so you decide to see if Rule B has any factual basis. To do this, you can investigate any of four locations (each card represents one location). Associated with each location is the egg cache of one bird species and the tracks of one mammal.

B:U-STD, U-SWC Story Summary: You are interested in whether Namka boys obey the laws of their people. Finding ostrich eggshells is a difficult, sophisticated task which represents a boy's transition into manhood. Duiker meat is a scarce and prized food; eating duiker meat at feasts is a privilege that must be earned, and is regulated by Rule B. You want to know if any of four boys at the feast cheated on this law when no one but you was looking (each card represents one boy).

Figure 6.5 Familiar Descriptive and Abstract Problems

Familiar Descriptive (F-D: Transportation Problem)

PAGE

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

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

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

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

A. subway

B. Arlington

C. cab

D. Boston

Abstract Problem (AP)

PAGE

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

"If a person has a 'D' rating, then his documents must be marked code '3'."

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

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

A. F

B. D

C. 3

D. 7

The instructions (reproduced in Figure 6.6) asked subjects to do the four tasks in order, without rereading any previous story or reviewing or changing any previous answers. The instructions were read aloud to subjects; in addition, subjects were given as much time as they wanted to read the instructions over to themselves before breaking the seal and beginning the experiment. Although most subjects completed the experiment in about 10 minutes, they were told they could take as much time as they wanted.

Results.

The percent of subjects choosing 'P & not-Q' for each problem closely matches the social contract predictions shown in Table 6.1. Hypotheses **AV** and **SC** both predict that the percent of

Table 6.2 Experiment 1: Percent of subjects choosing 'P & not-Q' or 'not-P & Q' for each problem (n=24)

	P not-Q	not-P Q
U-STD-SC:	75	0
U-D:	21	0
AP:	25	0
F-D:	46	0

Table 6.2 shows the percent of subjects who chose either 'P & not-Q' or 'not-P & Q'. Residual responses: These two categories do not, of course, exhaust all possible combinations of card choices -- there are sixteen in all. Of the 6 responses to the U-STD-SC that were not full SC answers, 5 were half correct, "sins of omission": 2 'P' responses (omitted not-Q) and 3 'not-Q' responses (omitted P). The 'not-Q' response is most interesting because not-Q is the card people routinely forget to choose on Wason selection tasks, and it is rarely chosen alone. If one counts 'not-Q' as also correct for both the U-STD-SC and the F-D, the magnitude of the difference between them increases (87.5% v. 54%).

Figure 6.6 Instructions

PLEASE DON'T LOOK THROUGH THIS BOOKLET YET
(But go ahead and read these instructions)

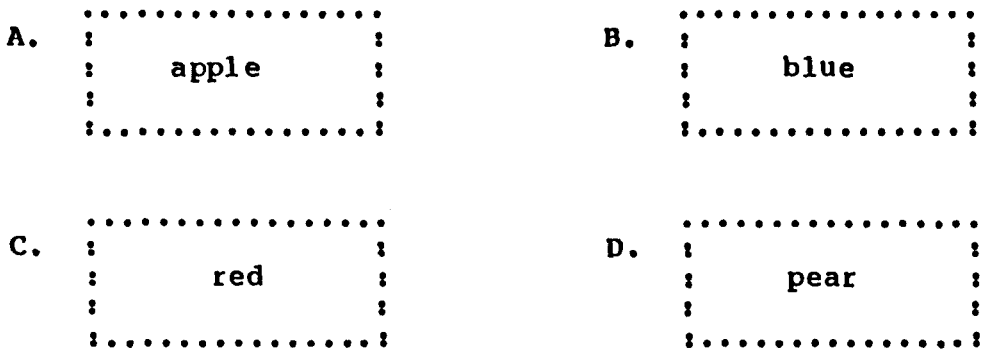
This is an experiment, completely optional, which will take about 15 minutes.

I am interested in how people think about different social situations. You will be reading some brief stories and then evaluating sentences like:

"If a fruit is an apple, then it must be red."

You will be evaluating the sentence with respect to information on four cards -- actually, pictures of four cards, like the pictures below.

In this example, each card would represent a fruit. Each card would have the name of a fruit on one side and that fruit's color on the other side, for example:



Of course, since they are only pictures, you will only be able to see one side of each card. The cards need not correspond to the way the world really is; for example, the "pear" card could say "yellow" or "red" or "purple" on the back, the "red" card could say "apple" or "banana" or "blueberry" on the back, and so on. For each story, you will be asked to indicate only those card(s) you definitely need to turn over to see if any of them violate the relation stated in the sentence. Circle the letter(s) (A,B,C, or D) which is next to the card(s) you want to turn over.

There are four different stories on four different pages. Don't look ahead at any stories: Read the first story and answer the question, then read the second story and answer its question, and so on, in the order the stories appear in this booklet. Please read a story in its entirety before you answer the question. Once you have finished answering the question associated with a story and gone on to the next story, do not go back and reread any previous stories or review or change any previous answers.

Never try to answer a question without first having read the entire story carefully. There are no "trick" questions. If a sentence or story seems ambiguous, use your common sense -- I am interested in what you would really do if faced with these situations in real life.

In sum: Pretend you really do have to investigate the situation described in a story. Then, for each card, ask yourself: "Would I need to see the information on the other side of this card in order to make a judgment?" If the answer is "yes", then circle the letter, A,B,C, or D, corresponding to that card.

Don't start until I have finished reading the instructions aloud. Take your time and have fun!

What year are you? Fresh ___, Soph ___, Jr ___, Sr ___, Other _____ (please specify)

Are you Female ___ or Male ___ ? Age _____ ?

subjects choosing 'not-P & Q' will be very low for these four problems; indeed, no one made this response in Experiment 1 (see Table 6.2).

Critical Tests

Predictions for critical tests 1 and 2 are taken from Table 6.1; for critical tests 3 and 4, they are taken from Table 6.3. Predictions for critical tests 5 and 6 are derived from both tables.

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

To answer this question, responses to the two unfamiliar problems must be compared; these problems use the same rule, but the story surrounding one rule makes it a social contract whereas the story surrounding the other makes it a descriptive rule.

Percentage 'P & not-Q' responses:

	Social Contract Prediction:	Availability Prediction:
U-STD-SC v. U-D.	U-STD-SC > U-D high low	U-STD-SC = U-D low low

75% > 21%

AV does not predict, and cannot account for, a wide discrepancy in falsifying ('P & not-Q') responses between these two unfamiliar problems. Yet a highly significant 54 point discrepancy occurred, just as **SC** predicts (75% v. 21%: $F_{1,23} = 27.18$, $p << .001$, $r = .74$*). U-STD-SC also produced a

* r is an effect size, which varies between zero and one (Rosenthal & Rosnow, 1984).

significant "content effect" when measured against the AP (75% v. 25%: $F_{1,23} = 13.80$, $p < .005$, $r = .61$). The U-D and AP both elicited the same low levels of falsifying responses (21% v. 25%: $F_{1,23} = 0.138$, n.s.).

Critical Test 2: Are there more SC responses to an <u>unfamiliar</u> standard social contract than falsifying responses to a <u>familiar</u> descriptive problem?

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

Percentage 'P & not-Q' responses:

	Social Contract Prediction:	Availability Prediction:	
U-STD-SC v. F-D	U-STD-SC > F-D	U-STD-SC ≤ F-D	
	high low*	low	mid-low
	75% > 46%		

As social contract theory predicts, an unfamiliar social contract (U-STD-SC) with which no subject could have had any actual experience elicited significantly more "falsifying" responses (SC responses) than a familiar relation (F-D) with which subjects were likely to have had experience (75% v. 46%: $F_{1,23} = 5.24$, $p <$ direction. Counting rare residuals (see legend, Table 6.2) for both problems magnifies U-STD-SC's advantage (87.5% v. 54%: $F_{1,23} = 5.41$, $p < .05$, $r = .44$).

* On Table 6.1, <u>SC</u> predictions regarding magnitudes for F-D problems assume no effect of availability; actually, <u>SC</u> is silent on whether or not availability exercises an independent effect on familiar problems.

Experiment 2

Experiment 2 was identical to Experiment 1, except the unfamiliar rules were switched rather than standard. The four problems fell into the following theoretical categories:

U-SWC-SC: Unfamiliar - Switched Social Contract
U-D: Unfamiliar - Descriptive
AP: Abstract Problem
F-D: Familiar - Descriptive

Table 6.3 shows the relative percentages of 'P & not-Q' and 'not-P & Q' responses expected in Experiment 2, assuming that responses are determined by either SC algorithms or availability, but not both.

Table 6.3 Predictions, Experiment 2: Social contract theory versus availability theory.

SWITCHED Social Contract (SWC-SC) v. Descriptive problems.

	'P & not-Q'		'not-P & Q'	
	Social Contract	Availability	Social Contract	Availability
U-SWC-SC:	very low	low	high	very low
U-D:	low	low	very low	very low
AP:	low	low	very low	very low
F-D:	low	middling to low	very low	very low

Table 6.3 Relative percentages of 'P & not-Q' and 'not-P & Q' responses expected for Experiment 2, assuming that responses are determined by either SC algorithms or availability, but not both.

'P & not-Q' responses: Social Contract Predictions: These are the cards SC algorithms should ignore on switched social contract problems; hence the percentage of "falsifying" responses should be very low on the U-SWC-SC. The rationale for the other predictions is the same as that presented in Table 6.1 for Experiment 1.

'not-P & Q' responses: Both hypotheses predict a very low percentage for all problems other than the U-SWC-SC, for which, according to social contract theory, it is the predicted response. See rationale in Table 6.1.

Subjects.

Twenty-four undergraduates from Harvard University participated in Experiment 2; they were paid volunteers recruited by advertisement (11 females, 13 males; mean age: 19.0 years).

Materials and Procedure.

The procedure was identical to that for Experiment 1. The materials were also identical with one exception: for unfamiliar rules, the propositions were switched. Thus, the 'A' version rule for the U-D and U-SWC-SC problems was: "If a man has a tattoo on his face then he eats cassava root,"* and the 'B' version rule was "If you have found an ostrich eggshell, then you eat duiker meat."

Results.

Table 6.4 shows the percent of subjects choosing 'P & not-Q' and 'not-P & Q' for each problem: these figures closely match the social contract predictions (see Table 6.3).

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

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

---

* "must" was left out of the "If" clause because it violates common English usage.

Table 6.4 Experiment 2: Percent of subjects choosing 'P & not-Q' or 'not-P & Q' for each problem (n=24)

	P not-Q	not-P Q
U-SWC-SC:	4	67
U-D:	12	4
AP:	12	0
F-D	50	0

Table 6.4 shows the percent of subjects who chose either 'P & not-Q' or 'not-P & Q' in Experiment 2. Residual responses: Of the 8 responses to the U-SWC-SC that were not full SC answers, 6 were "sins of omission": 1 'not-P' (omitted Q) and 5 'Q' (omitted not-P). Both are rare answers on Wason selection tasks, and the frequency of 'Q' responses is much higher than its expected value based on the other problems in Exps 1 and 2 ($Z=4.00$, $p<.00003$). If one counts rare, half-correct residuals as also correct ('Q' for U-SWC-SC, 'not-Q' for F-D), the magnitude of the difference between U-SWC-SC and F-D increases (87.5% v. 54%), exactly matching the U-STD-SC v. F-D figures in Exp 1 for the equivalent categorization scheme.

Percentage 'not-P & Q' responses:

	Social Contract Prediction:	Availability Prediction:
U-SWC-SC v. U-D,AP,F-D.	U-SWC-SC > U-D,AP,F-D high very low	U-SWC-SC = U-D,AP,F-D very low very low
	67% > 4%, 0%, 0%	

The large and significant 63-67 point difference between the U-SWC-SC and all other problems (67% v. 4%, 0%, 0%, L = +3,-1,-1,-1: $F_{1,69} = 116.26$, $p << .001$, $r = .79$) is predicted only by SC. 'Not-P & Q' was chosen only once on any non-SWC-SC problem in Experiment 2, and was not chosen by anyone in Experiment 1.

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

As in Critical Test 2, one can use the percentage of falsifying responses to the F-D problem as a metric for judging the size of the social contract effect. This requires that the proportion of SC responses ('not-P & Q') to the U-SWC-SC be compared to the proportion of falsifying responses to the F-D.

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

	Social Contract Prediction:	Availability Prediction:
U-SWC-SC v. F-D.	U-SWC-SC > F-D high low	U-SWC-SC < F-D very low mid-low
	67% > 50%	

As SC predicts, SC responses to the U-SWC-SC outstripped falsifying responses to the F-D (67% v. 50%: $F_{1,23} = 2.09$, n.s.). The difference is not significant; however AV predicts an inequality in the opposite direction. When rare residuals are counted for both problems (U-SWC-SC: 'Q'; F-D: 'not-Q'), the difference is magnified to the exact proportions found for the parallel U-STD-SC v. F-D comparison in Experiment 1, and is significant (87.5% v. 54%: $F_{1,23} = 8.36$, $p < .01$, $r = .52$). This supports the contention that SC algorithms are a major determinant of responses to problems involving social exchange, even when those problems are unfamiliar.

Experiment 1 versus Experiment 2

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

Because the U-STD-SC and U-SWC-SC are both unfamiliar problems, **AV** predicts they should both elicit low levels of 'P & not-Q' responses. The social contract prediction could not be more different. For a STD-SC, P represents the "benefit accepted" card and not-Q represents the "cost NOT paid" card, the cards that a "look for cheaters" procedure should choose. However, for a SWC-SC, P represents the "cost paid" card and not-Q represents the "benefit NOT accepted" card, the cards a "look for cheaters" procedure should ignore because they represent people who could not possibly have cheated (see Figure 6.1). What logical category these cards fall into is simply irrelevant from a social contract perspective. Cards should be chosen on the basis of their cost/benefit category, not their logical category.

Percentage 'P & not-Q' responses:

	Social Contract Prediction:	Availability Prediction:
U-STD-SC v. U-SWC-SC.	U-STD-SC >> U-SWC-SC high very low 75% >> 4%	U-STD-SC = U-SWC-SC low low

The large and significant 71 point discrepancy in 'P & not-Q' responses between U-STD-SC and U-SWC-SC problems is predicted only by **SC** (75% v. 4%: $Z = 5.02$, $p < .0000005$, phi = .72*). Furthermore, the **SC** prediction that the dominant, SC response to

the U-STD-SC will be very rare on the U-SWC-SC was borne out. Only one subject gave the STD-SC answer, 'P & not-Q', in response to the U-SWC-SC -- and this was one of only two subjects in Exp 2 to give falsifying answers to all of the three other problems.

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

Critical Test 6 is simply the flip side of Critical Test 5. The predicted SC response to a SWC-SC is 'not-P & Q': not-P represents the "cost NOT paid" card and Q represents the "benefit accepted" card (see Figure 6.1). But for a STD-SC, not-P represents the "benefit NOT accepted" card and Q represents the "cost paid" card -- the cards a "look for cheaters" procedure should ignore, regardless of their logical category. Hence, SC predicts that the correct SC answer to a SWC-SC, 'not-P & Q', will be very rare for a STD-SC. In contrast, AV predicts that the percentage of subjects choosing 'not-P & Q' on the U-STD-SC and the U-SWC-SC will be about equal, and very low (see Table 6.1).

Percentage 'not-P & Q' responses:

	Social Contract Prediction:		Availability Prediction:	
U-SWC-SC v. U-STD-SC.	U-SWC-SC high	>> U-STD-SC very low	U-SWC-SC very low	= U-STD-SC very low
	67%	>> 0%		

The large and significant 67 point discrepancy in 'not-P & Q' responses between U-SWC-SC and U-STD-SC problems is predicted only by SC (67% v. 0%: $Z = 4.90$, $p < .0000005$, phi = .71) .

* phi is an effect size, which varies between zero and one (Rosenthal & Rosnow, 1984).

Furthermore, the **SC** prediction that the dominant, SC response to the U-SWC-SC will be very rare on the U-STD-SC was borne out: no one gave the SWC-SC answer, 'not-P & Q', in response to the U-STD-SC.

Summary, Critical Tests for Experiments 1 and 2.

Because the **AV** and **SC** hypotheses make very different predictions regarding six comparisons between problems, critical tests between these two hypotheses can be constructed from the predictions of Table 6.1. The results for each of the six tests verify the **SC** prediction and falsify the **AV** prediction (see Figure 6.7 for summary of critical tests and results).

Social Contract Tests

Critical tests only address the question: Is the data better explained by social contract theory or availability theory? However, there are other questions one can ask of this data that are specific to social contract theory. Because these questions involve a comparison of results from Experiments 1 and 2, for convenience, the data from Tables 6.2 and 6.4 are combined into Table 6.5 below.

Are the logically distinct SC answers to standard and switched SC problems produced by the same algorithms?

The correct SC answers for standard and switched social contracts are very different from a logical point of view: 'P & not-Q' for U-STD-SC v. 'not-P & Q' for U-SWC-SC. However, the proportions of SC answers to the U-STD-SC (75%) and U-SWC-SC (67%) are not significantly different ($Z = 0.63$), just as one

Figure 6.7

CRITICAL TESTS: SOCIAL CONTRACT THEORY VERSUS AVAILABILITY THEORY

	CRITICAL TEST:	SOCIAL CONTRACT PREDICTION (*SC*):	AVAILABILITY PREDICTION (*AV*):	WHICH HYPOTHESIS DO THE DATA SUPPORT?	CONCLUSION:
EXPERIMENT 1 CRITICAL COMPARISONS	*CRITICAL TEST 1: Does an unfamiliar STANDARD social contract elicit the predicted SC response, 'P & not-Q'?*				
	percentage '*P & not-Q*' responses: **1. U-STD-SC v. U-D**	**U-STD-SC > U-D** high low	**U-STD-SC = U-D** low low	***SC*: 75% > 21%,** $F_{1,23}$=27.18, p<<.001, r =.74*	***SC* verified,** ***AV* falsified.**
	CRITICAL TEST 2: Are there more SC responses to a standard UNFAMILIAR social contract problem than falsifying responses to a FAMILIAR descriptive problem?				
	percentage '*P & not-Q*' responses: **2a. U-STD-SC v. F-D**	**U-STD-SC > F-D** high low	**U-STD-SC ≤ F-D** low mid-low	***SC*: 75% > 46%,** $F_{1,23}$=5.24, p<.05, r =.43	***SC* verified,** ***AV* falsified.**
	plus rare residuals ('*not-Q*' for both): **2b. U-STD-SC v. F-D**	**same as 2a**	**same as 2a**	***SC*: 87.5% > 54%,** $F_{1,23}$=5.41, p<.05, r =.44	
EXPERIMENT 2 CRITICAL COMPARISONS	*CRITICAL TEST 3: Does an unfamiliar SWITCHED social contract elicit the predicted SC response, 'not-P & Q'?*				
	percentage '*not-P & Q*' responses: **3. U-SWC-SC v. U-D,AP,F-D**	**U-SWC-SC > U-D,AP,F-D** high v. low	**U-SWC-SC = U-D,AP,F-D** v. low v. low	***SC*: 67% > 4%, 0%, 0%** *L*=+3,-1,-1,-1: $F_{1,69}$=116.26, p<<.001, r =.79	***SC* verified,** ***AV* falsified.**
	CRITICAL TEST 4: Are there more SC responses to a switched UNFAMILIAR social contract problem than falsifying responses to a FAMILIAR descriptive problem?				
	percentage '*not-P & Q*' responses to U-SWC-SC; percentage '*P & not-Q*' to F-D: **4a. U-SWC-SC v. F-D**	**U-SWC-SC > F-D** high low	**U-SWC-SC < F-D** v. low mid-low	***SC*: 67% > 50%** correct direction, but n.s. ($F_{1,23}$=2.09)	***SC* verified,** ***AV* falsified.**
	plus rare residuals (U-SWC-SC: '*Q*'; F-D: '*not-Q*'): **4b. U-SWC-SC v. F-D**	**same as 4a**	**same as 4a**	***SC*: 87.5% > 54%,** $F_{1,23}$=8.36, p<.01, r =.52	
EXP 1 v. EXP 2 CRITICAL COMPARISONS	*CRITICAL TESTS 5 and 6: Is the correct SC response to a STANDARD social contract very rare for a SWITCHED social contract, and vice versa?*				
	percentage '*P & not-Q*' responses: **5. U-STD-SC v. U-SWC-SC**	**U-STD-SC >> U-SWC-SC** high v. low	**U-STD-SC = U-SWC-SC** low low	***SC*: 75% >> 4%,** Z=5.02, p<.0000005, phi=.72*	***SC* verified,** ***AV* falsified.**
	percentage '*not-P & Q*' responses: **6. U-SWC-SC v. U-STD-SC**	**U-SWC-SC >> U-STD-SC** high v. low	**U-SWC-SC = U-STD-SC** v. low v. low	***SC*: 67% >> 0%,** Z=4.90, p<.0000005, phi=.71	***SC* verified,** ***AV* falsified.**

AVAILABILITY ASSESSED: EXPERIMENTS 1 AND 2

EXP 1 &. EXP 2	*AVAILABILITY TEST: Does availability have any effect at all on familiar problems? (standard test in literature: F-D > AP)*				
	percentage '*P & not-Q*' responses: **7. F-D v. AP, U-D**	not applicable	**F-D > AP, U-D** mid-low low	**Exp 1: 46% > 25%, 21%** **Exp 2: 50% > 12%, 12%**	**Availability has a marginal effect on familiar problems**
	Exp 1: F-D v. AP: NO EFFECT ($F_{1,23}$=4.02, n.s.); **F-D v. U-D: EFFECT** ($F_{1,23}$=5.31, p<.05) **Exp 2: F-D v. AP: EFFECT** ($F_{1,23}$=13.80, p<.005); **F-D v. U-D: EFFECT** ($F_{1,23}$=13.80, p<.005)				

Table 6.5 Experiments 1 and 2: Percent of subjects choosing 'P & not-Q' or 'not-P & Q' for each problem

	Experiment 1 (n=24)				Experiment 2 (n=24)			
	P not-Q		not-P Q			P not-Q		not-P Q
U-STD-SC:	75*		0		U-SWC-SC:	4		67*
U-D:	21		0		U-D:	12		4
AP:	25		0		AP:	12		0
F-D:	46		0		F-D	50		0

*predicted SC response to social contract problems.

would expect if the same SC algorithms were producing these two, logically distinct, responses. When rare residuals (see legend, Tables 6.2 and 6.4) are added in, the percentage of SC answers on these two problems is identical -- 87.5%. Using percent falsifying answers to the U-D (same rule as U-SC) as a baseline for comparison, the relative advantage SC status gave in producing SC answers is almost identical for both SC problems: 54 points between U-STD-SC and its U-D, 55 points between U-SWC-SC and its U-D.

Table 6.6 shows the frequencies with which individual cards were selected in Experiments 1 and 2.

Table 6.6 Experiments 1 and 2: Selection frequencies for individual cards, sorted by logical category and social contract category

Logical Category:	U-D		AP		F-D		U-SC STD	U-SC SWC	Social Contract Category:	U-SC STD	U-SC SWC
	Exp 1	Exp 2	Exp 1	Exp 2	Exp 1	Exp 2	Exp 1	Exp 2		Exp 1	Exp 2
P	21	19	23	23	22	22	20	2	Benefit Accepted	20	21
not-P	5	6	6	6	1	1	1	17	Benefit NOT Accepted	1	1
Q	9	12	8	14	4	1	0	21	Cost Paid	0	2
not-Q	6	7	11	11	13	14	22	1	Cost NOT Paid	22	17

When cards are sorted according to their logical category, all

problems replicate nicely over Experiments 1 and 2, except the social contract problems. When sorted according to logical category, selection frequencies for the U-STD-SC and U-SWC-SC are radically at variance with one another. When sorted according to social contract category, however, their profiles are almost identical. This indicates that for unfamiliar social contract problems, a social contract categorization scheme captures dimensions that are psychologically real for subjects, whereas a logical categorization scheme does not.

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

If SC algorithms are, in part, frame-builders, as proposed, one would expect them to operate in novel social exchanges, as well as in familiar ones. In fact, the 75% "falsification" rate on the U-STD-SC, an unfamiliar social contract, is equivalent to that usually found for the Drinking Age Problem, a highly familiar standard social contact (Cox & Griggs, 1982; Griggs & Cox, 1982; Griggs & Cox, 1983; see Chapter 2). This cannot be accounted for by differences in subject populations: 78% of a similar group of 23 Harvard undergraduates "falsified" on the DAP (see Experiment 6-A below). Thus, the percent of Harvard undergraduates choosing 'P & not-Q' on a STD-SC was the same, regardless of whether the social contract was very familiar or completely unfamiliar (78% v. 75%: $Z = 0.26$, n.s.).

The same was true of the unfamiliar switched social contract. The percentage of subjects choosing 'not-P & Q' on the U-SWC-SC did not differ significantly from the percentage choosing 'P & not-Q' for the familiar DAP (67% v. 78%: $Z = 0.88$,

n.s.). The hypothesis that unfamiliar SC problems generate fewer SC answers than familiar SC problems is not supported even if one uses all three problems (F-STD-SC, U-STD-SC, U-SWC-SC) in one test (78% v. 75%, 67%, L = +2,-1,-1, $F_{1,68}$ = 0.43, n.s.). The data of Experiments 1 and 2 suggest that SC algorithms work just as well in novel social exchanges as in familiar ones -- just as frame-builders should.

Availability Assessed

Does availability have any effect at all on familiar problems?

Although AV cannot explain the precisely patterned differences in performance among unfamiliar social contracts and all other problems, availability does appear to have had a minor, somewhat erratic effect on familiar descriptive problems (see Figure 6.7).

In the literature, a standard test of the efficacy of availability is to compare an F-D to an AP. The standard "now you see it, now you don't" result of such experiments (see Chapter 2) is mirrored in Experiments 1 and 2. The difference in percentage of falsifying responses between the F-D and AP is significant in Experiment 2, but not in Experiment 1. However, the F-D does fare significantly better than the U-D in both experiments. Furthermore, the set of contrasts for Experiment 1 which assumes that the F-D outstrips both AP and U-D problems (L=+2,-1,-1) is significant ($F_{1,46}$ = 5.97, $p < .025$).

These results indicate that availability can give familiar non-SC problems an advantage of 21 to 38 points over unfamiliar

non-SC problems in producing falsifying responses (average advantage = 30.5). However, using percent falsifying responses to U-Ds and APs as a baseline for comparison, the availability advantage appears to be much less important than the social contract effect. The average advantage in producing SC responses that social contract status gives to an unfamiliar problem is about 1.8 times the size of the availability advantage. And, as Critical Tests 2 and 4 of Figure 6.7 show, more SC responses were elicited by unfamiliar social contract problems than falsifying responses by familiar non-SC problems that had an availability advantage.

Summary, Experiments 1 and 2.

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

Experiment 3

In Experiments 1 and 2, the unfamiliar social contracts used were expressed as laws of one's social group. I did this because the rules used in the literature on the Wason selection task were invariably expressed as laws. Moreover, using a social contract law let me use the exact same rule in U-SC and U-D problems.

However, social contract algorithms should work just as well with conditionals that express a private social exchange between just two individuals, rules like:

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

or, equivalently,

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

To test this, I conducted two experiments (Experiments 3 and 4) that were identical to Experiments 1 and 2, except the unfamiliar social contracts used expressed an exchange between two individuals rather than a social law.

The predictions for Experiments 3 and 4 are identical to the predictions for Experiments 1 and 2. Hence, they provide an opportunity to replicate the results of the six critical tests for choosing between social contract theory and availability theory.

Subjects.

Twenty-four undergraduates from Harvard University participated in Experiment 3; they were paid volunteers, recruited by advertisement (11 females, 13 males; mean age: 20.0 years (no data on age of 3 subjects)).

Materials and Procedures.

The procedure was identical to that described for Experiment 1. The materials were also identical, with one exception: the U-STD-SC expressed a private exchange rather than a social law. The 'A' version rule was: "If you get a tattoo on your face, then I'll give you cassava root." The 'B' version rule was: "If you give me your ostrich eggshell, then I'll give you duiker meat." In both cases, the "deal" was offered by the person identified in the story as the potential cheater. Thus, in terms of the value

system of the potential cheater, the SC structure of both problems was: "If B(me) then C(me)" -- a STD-SC. The SC answer to such a problem is 'P & not-Q'. The U-STD-SC problems used in Experiment 3 are shown in Figure 6.8.

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

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

Figure 6.8 Experiment 5: 'A' and 'B' versions

' version U-STD-SC

PAGE

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

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

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

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

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

You learn that Big Kiku hates some of these men for betraying him to his enemies. You suspect he will cheat and betray some of them. Thus, this is a perfect opportunity for you to see first hand how Big Kiku wields his power. The cards below have information about the fates of the four men. Each card represents one man. One side of a card tells whether or not the man went through with the facial tattoo that evening and the other side of the card tells whether or not Big Kiku gave that man cassava root the next day.

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

A.	got the tattoo	B.	Big Kiku gave him nothing
C.	no tattoo	D.	Big Kiku gave him cassava root

'B' version U-STD-SC

PAGE

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

Bo is a crafty old Namka man in the band you are studying. He is always accidentally breaking his ostrich eggshell and would like to "stockpile" some -- the Namka use ostrich eggshells as canteens because they are light and hold lots of water. He sees his opportunity when four men from a neighboring band stumble into camp one morning.

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

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

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

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

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

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

A.	He gave his ostrich eggshell to Bo	B.	Bo gave him nothing
C.	He gave Bo nothing	D.	Bo gave him duiker meat

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

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

Because the social contract rules in Experiments 1 and 2 were expressed as social laws, the U-SC rules could be identical to their corresponding U-D rules. In experiments 3 and 4, the terms used are similar to those used in the corresponding U-D rules, but, because the U-SC rules express private deals rather than laws, they could not be identical.

U-STD-SC: "If you get a tattoo on your face, then I'll give you cassava root."

U-D: "If a man has a tattoo on his face, then he eats cassava root."

U-STD-SC: "If you give me your ostrich eggshell, then I'll give you duiker meat."

U-D: "If you have found an ostrich eggshell, then you eat duiker meat."

Results.

The percent of subjects choosing 'P & not-Q' for each problem closely matches the social contract predictions shown in Table 6.1. No one chose 'not-P & Q'; this is consistent with both AV and SC. The results are remarkably similar to those for

Experiment 1.

Table 6.7 Experiment 3: Percent of subjects choosing 'P & not-Q' or 'not-P & Q' for each problem (n=24)

	P not-Q	not-P Q
U-STD-SC:	71	0
U-D:	25	0
AP:	29	0
F-D:	38	0

Table 6.7 shows the percent of subjects who chose either 'P & not-Q' or 'not-P & Q'. Residual responses: Of the 7 responses to the U-STD-SC that were not full SC answers, 5 were half correct, "sins of omission": 5 'P' responses (omitted not-Q). No one answered 'not-Q' (omitting P) on either the U-STD-SC or the F-D, so counting rare residuals neither increases nor decreases the magnitude of the difference between these two problems.

Critical Tests

Six critical tests pitting hypothesis **AV** against hypothesis **SC** were used in analyzing the data from Experiments 1 and 2. The same six critical tests can be carried out on the data from Experiments 3 and 4.

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

Percentage 'P & not-Q' responses:

	Social Contract Prediction:	Availability Prediction:
U-STD-SC v. U-D.	U-STD-SC > U-D high low	U-STD-SC = U-D low low

71% > 25%

There is a highly significant 46 point discrepancy in 'P & not-Q' responses between the U-STD-SC and the U-D (71% v. 25%: $F_{1,23}$ = 19.46, $p < .001$, $r = .68$). This discrepancy is predicted only by SC; AV predicts a low proportion of falsifying responses on all unfamiliar problems, whether they are social contracts or not. U-STD-SC also produces a significant "content effect" when measured against the AP (71% v. 29%: $F_{1,23}$ = 16.43, $p < .001$, $r = .65$). The U-D and AP both elicited the same low levels of falsifying responses (25% v. 29%: $F_{1,23} = 0.19$, n.s.).

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

Percentage 'P & not-Q' responses:

	Social Contract Prediction:	Availability Prediction:
U-STD-SC v. F-D	U-STD-SC > F-D	U-STD-SC ≤ F-D
	high low	low mid-low
	71% > 38%	

The advantage that SC status gives an unfamiliar problem is larger than the advantage availability gives a familiar descriptive problem. The U-STD-SC elicited significantly more "falsifying" (SC) responses than the F-D (71% v. 38%: $F_{1,23}$ = 6.57, $p < .025$, $r = .47$). AV predicts an inequality in the opposite direction. This supports the claim that SC algorithms are a major determinant of responses to problems involving social exchange.

Experiment 4

This experiment is identical to Experiment 2, except the

switched social contract rules used express a private exchange rather than a social law. AV and SC predictions are the same as for Experiment 2.

Subjects.

Twenty-four undergraduates from Harvard University participated in Experiment 4; they were paid volunteers, recruited by advertisement (11 females, 13 males; mean age: 19.4 years (no data on one subject's age)).

Materials and Procedures.

The procedure was identical to that for Experiment 3. The materials were also identical with one exception: the unfamiliar rules were "switched." Thus, the U-SWC-SC rules were: "If I give you cassava root, then you must get a tattoo on your face" ('A' version), and "If I give you duiker meat, then you must give me your ostrich eggshell" ('B' version). The U-Ds were the same as in Experiment 2.

Results.

The results are shown in Table 6.8; they match the social contract predictions of Table 6.3.

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

Percentage 'not-P & Q' responses:

	Social Contract Prediction:	Availability Prediction:
U-SWC-SC v. U-D,AP,F-D.	U-SWC-SC > U-D,AP,F-D high very low	U-SWC-SC = U-D,AP,F-D very low very low
	75% > 0%, 0%, 0%	

Table 6.8 Experiment 4: Percent of subjects choosing 'P & not-Q' or 'not-P & Q' for each problem (n=24)

	P not-Q	not-P Q
U-SWC-SC:	0	75
U-D:	25	0
AP:	33	0
F-D:	58	0

Table 6.8 shows the percent of subjects who chose either 'P & not-Q' or 'not-P & Q'. Residual responses: Of the 6 responses to the U-SWC-SC that were not full SC answers, 4 were half correct, "sins of omission": 4 'Q' responses (omitted not-P). Counting rare residuals for both the U-SWC-SC ('Q') and the F-D ('not-Q') increases the magnitude of the difference between them (92% v. 63%).

The large and significant 75 point difference between U-SWC-SC and all other problems is predicted only by SC (75% v. 0%, 0%, 0%, L = +3,-1,-1,-1: $F_{1,69}$ = 207.01, $p << .001$, $r = .87$). 'Not-P & Q' was not chosen on any other problem in both Experiments 3 and 4. AV can neither predict nor explain this result.

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

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

	Social Contract Prediction:	Availability Prediction:
U-SWC-SC v. F-D.	U-SWC-SC > F-D high low	U-SWC-SC < F-D very low mid-low
	75% > 58%	

There were more SC responses to the unfamiliar U-SWC-SC than

falsifying responses to the familiar F-D, just as SC predicts (75% v. 58%: $F_{1,23}$ = 1.64, n.s.). Although the difference is not significant, AV predicts an inequality in the opposite direction. When rare residuals are counted for both problems (U-SWC-SC: 'Q'; F-D: 'not-Q'), the difference is magnified, and is significant (92% v. 63%: $F_{1,23}$ = 6.75, $p < .025$, $r = .48$). Like the results of Critical Test 2, this supports the contention that SC algorithms are a major determinant of responses to problems involving social exchange.

Experiment 3 versus Experiment 4

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

AV predicts that the U-STD-SC and the U-SWC-SC should both elicit low levels of 'P & not-Q' responses because they are both unfamiliar. In contrast, SC predicts that 'P & not-Q', the dominant, SC response to the U-STD-SC, should be very rare on the U-SWC-SC.

Percentage 'P & not-Q' responses:

	Social Contract Prediction:	Availability Prediction:
U-STD-SC v. U-SWC-SC.	U-STD-SC >> U-SWC-SC high very low	U-STD-SC = U-SWC-SC low low
	71% >> 0%	

Only SC predicts the large and significant 71 point discrepancy in 'P & not-Q' responses between the U-STD-SC and the U-SWC-SC (71% v. 0%: $Z = 5.14$, $p < .00000025$, phi = .74). The SC

prediction that the dominant, SC response to the U-STD-SC will be very rare on the U-SWC-SC was also borne out: no one gave the STD-SC answer, 'P & not-Q', in response to the U-SWC-SC.

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

SC predicts that the correct SC answer to a SWC-SC, 'not-P & Q', will be very rare for a STD-SC. In contrast, AV predicts that the percentage of subjects choosing 'not-P & Q' on the U-STD-SC and the U-SWC-SC will be about equal, and very low.

Percentage 'not-P & Q' responses:

	Social Contract Prediction:		Availability Prediction:	
U-SWC-SC v. U-STD-SC.	U-SWC-SC >> high	U-STD-SC very low	U-SWC-SC = very low	U-STD-SC very low
	75% >> 0%			

The large and significant 75 point discrepancy in 'not-P & Q' responses between U-SWC-SC and U-STD-SC problems is predicted only by SC (75% v. 0%: $Z = 5.37$, $p < .0000001$, phi = .77). Furthermore, the dominant, SC response to the U-SWC-SC was indeed very rare on the U-STD-SC, just as SC predicts: no one gave the SWC-SC answer, 'not-P & Q', in response to the U-STD-SC.

Social Contract Tests

As before, if the social contract view is correct, certain relations should be manifest in the data, above and beyond those addressed by the critical tests. For convenience, the data from Tables 6.7 and 6.8 are combined in Table 6.9.

Table 6.9 Percent of subjects choosing 'P & not-Q' or 'not-P & Q' for each problem

	Experiment 3 (n=24)				Experiment 4 (n=24)		
	P not-Q	not-P Q			P not-Q	not-P Q	
U-STD-SC:	71*	0		U-SWC-SC:	0	75*	
U-D:	25	0		U-D:	25	0	
AP:	29	0		AP:	33	0	
F-D:	38	0		F-D	58	0	

*predicted SC response to social contract problems.

Are the logically distinct SC answers to standard and switched SC problems produced by the same algorithms?

The proportions of SC answers to the U-STD-SC ('P & not-Q') and U-SWC-SC ('not-P & Q') are not significantly different (71% v. 75%: Z = 0.32, n.s.), just as one would expect if these two logically distinct responses were the product of the same SC algorithms. Using percent falsifying answers to the U-D (same rule as U-SC) as a baseline for comparison, the relative advantage SC status gave in producing SC answers is very similar for both SC problems: 46 points between U-STD-SC and its U-D, 50 points between U-SWC-SC and its U-D.

For these personal exchange problems, residual responses did not split (e.g., some 'P', some 'not-Q' for the U-STD-SC) as they did for the law problems of Experiments 1 and 2; the only "sins of omission" in Experiments 3 and 4 involved choosing the card that represents what the honest person did. There were 5 'P'

alone responses for the U-STD-SC and 4 'Q' alone responses to the U-SWC-SC. This is interesting, because only Experiments 3 and 4 admit the possibility of intercontingent behavior; these are the responses one would expect if a subject read through quickly, not noticing that there is a time delay between when the potential cheater gets his benefit and when he is expected to honor his end of the deal. The numbers involved are too small to firmly attribute this pattern to the power of intercontingent reasoning, but it is an area that deserves future research.

Table 6.10 shows the frequencies with which individual cards were selected in Experiments 3 and 4.

Table 6.10 Experiments 3 and 4: Selection frequencies for individual cards, sorted by logical category and social contract category

Logical Category:	U-D		AP		F-D		U-SC STD	U-SC SWC	Social Contract Category:	U-SC STD	U-SC SWC
	Exp 3	Exp 4	Exp 3	Exp 4	Exp 3	Exp 4	Exp 3	Exp 4		Exp 3	Exp 4
P	22	21	23	23	24	23	23	1	Benefit Accepted	23	23
not-P	4	5	4	5	0	1	2	19	Benefit NOT Accepted	2	2
Q	11	11	7	9	4	5	0	23	Cost Paid	0	1
not-Q	12	13	10	14	10	18	17	2	Cost NOT Paid	17	19

Just as before, the two social contract problems replicate when sorted by social contract category, but not when sorted by logical category, as the other problems do. The social contract categorization scheme captures dimensions that are psychologically real for the subjects in Experiments 3 and 4, just as it did in Experiments 1 and 2.

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

A frame-builder structures novel experiences along evolutionarily relevant dimensions that it is keyed to pick up. If SC algorithms function as frame-builders, then they should operate well in novel social exchanges, like those represented in

the unfamiliar social contract problems. The percentage of SC responses elicited by the U-STD-SC -- 71% -- does not differ significantly from the 78% elicited by the familiar DAP with a similar group of Harvard students (71% v. 78%: Z = 0.58, n.s.; see Experiment 6-A). Neither does the percentage of logically distinct SC responses elicited by the U-SWC-SC (75% v. 78%: Z = 0.26, n.s.). The hypothesis that familiar SC problems elicit more SC answers than unfamiliar ones is not supported by this data, even when one combines all three problems (F-STD-SC, U-STD-SC, U-SWC-SC) into one test (78% v. 71%: 75%, L = +2,-1,-1, $F_{1,68}$ = 0.23, n.s.). Not only are SC algorithms keyed to abstract SC dimensions from novel situations, but they accomplish this as efficiently in novel situations as they do in familiar ones.

Availability assessed

Does availability have any effect at all on familiar problems?.

The F-D did not elicit significantly more falsifying responses to either the AP or the U-D in Experiment 3 (F-D v. AP: 38% v. 29%, $F_{1,23}$ = 1.31, n.s.; F-D v. U-D: 38% v. 25%, $F_{1,23}$ = 1.30, n.s.; F-D v. AP, U-D: L = +2,-1,-1, $F_{1,46}$ = 1.38, n.s.). However, in Experiment 4 the F-D did elicit more falsifying responses than both the AP and U-D (F-D v. AP: 58% v. 33%, $F_{1,23}$ = 5.31, $p < .05$, $r = .43$; F-D v. U-D: 58% v. 25%, $F_{1,23}$ = 11.50, $p <$ these two experiments than it was in Experiments 1 and 2.

In Experiments 3 and 4, availability gave familiar non-SC problems an advantage of 9 to 33 points over unfamiliar non-SC problems in producing falsifying responses (average advantage =

20). However, social contract status gave unfamiliar problems an average advantage of 45 points in producing SC responses -- this means the social contract advantage was 2.25 times the size of the availability advantage. Compared to the social contract effect, the effect of availability was even smaller in Experiments 3 and 4 than it was in Experiments 1 and 2.

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

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

Summary, Experiments 3 and 4.

These experiments replicated the results of Experiments 1 and 2: as before, the six critical tests supported hypothesis **SC** over **AV**, and all the other social contract predictions were borne

out. Availability was shown, once again, to have a minor but erratic effect on familiar descriptive problems.

Position Effects, Experiments 1 through 4

Experiments 1 through 4 controlled for problem position: each experiment had four problems, and every problem occurred in each position the same number of times. Therefore these experiments can be used to see if there were any position effects: if the probability of a subject solving a Wason selection task correctly increases with the number of problems that subject has already been exposed to. This will be useful to know in analyzing some data from Experiment 6.

Experiments 1 through 4 do not allow a pure test of position,* however, because the possibility of transfer from a successfully solved social contract problem is confounded with position -- the later a problem occurs in the booklet, the more likely it is to have followed the social contract problem. Although no one in the literature has yet found transfer from a successfully solved Wason selection task to an abstract problem (Johnson-Laird, Legrenzi, & Legrenzi, 1972; Griggs & Cox, 1982), there is one report of transfer from a social contract problem to a "semi" social contract problem, the apparel-color problem (Cox & Griggs, 1982; see Chapter 5). Success on a social contract problem could give subjects insight into the structure of the F-D transportation problem, especially in my subject population. After all, my subject population is an unusual one: the median SAT scores of Harvard students are among the highest -- if not the highest -- in the U.S. SATs test, in part, an applicants'

ability to discern the structure of a problem, regardless of its content; high scores indicate skilled problem solvers. Consonant with this, for my subject population, the falsification rate on the abstract problems of Experiments 1-4 was unusually high: instead of the 4-10% rate found for most studies (Wason, 1983), an average of 25% of Harvard students falsified on the abstract problems. First solving a U-SC correctly, combined with the familiarity of the transportation problem, might provide enough insight into the structure of these problems to improve performance on the transportation problem.

Table 6.11 shows the number of problems correctly solved as a function of position for Experiments 1 through 4. The hypothesis that there is a linear effect of position was tested using the contrasts L = -3, -1, +1, +3.

Table 6.11 Number of problems correctly solved as a function of position.**

Position:		1	2	3	4	
Experiment:						
1	(n=24)	7	9	11	13	$F_{1,69}$ = 3.92, n.s.
2	(n=24)	8	7	8	11	$F_{1,69}$ = 1.02, n.s.
3	(n=24)	9	10	10	10	$F_{1,69}$ = 0.10, n.s.
4	(n=24)	8	12	13	13	$F_{1,69}$ = 3.12, n.s.
1-4	(n=96)	32	38	42	47	$F_{1,285}$ = 6.60, $p < .025$, $r = .15$
2-4	(n=72)	25	29	31	34	$F_{1,213}$ = 3.18, n.s.

** "Correct" is defined as 'not-P & Q' for U-SWC-SC problems, 'P & not-Q' for all other problems.

* It is not clear to me that a pure test of order is possible -- all methods I have considered seem riddled with confounds.

None of these experiments showed any significant position effect. When the results of all four experiments are combined into one large experiment (n=96) there is a significant effect of order. But remember: position is confounded with transfer in these experiments, if there is evidence of transfer. Experiment 1 showed significant transfer from a successfully solved U-SC to the F-D; none of the other experiments did (Exp 1: G(adj) = 6.34, $p < .01$; Exp 2: G(adj) = 0.17, n.s.; Exp 3: G(adj) = 0.00, n.s., Exp 4: G(adj) = 1.56, n.s.; Sokal & Rohlf, 1969, p.591). When Experiment 1 is excluded from the analysis such that the three experiments showing no evidence of transfer (2-4) are analyzed together, there is no significant order effect -- and, for an effect size of $r = .15$, 72 subjects are sufficient to produce a significant F. Thus, it seems fair to conclude that the order effect found for the combined 1-4 data was due to the transfer effect in Experiment 1, not to an effect of problem position. Taken together, these results indicate that there is no effect of position alone on performance.

Experiment 5

For standard social contracts, the correct SC response happens to also be the logically correct response, 'P & not-Q'. Experiments 1 and 3 showed that STD-SC status facilitates "logical falsification" for unfamiliar rules set using concrete -- if obscure -- terms. But can STD-SC status also facilitate "logical falsification" for an abstract problem?

Experiment 5 compared performance on two abstract, prescriptive rules: one a standard social contract rule, one not

Both were set in a realistic context -- they were bureaucratic rules involving an immigration office. The story surrounding one rule defined its propositions in cost/ benefit terms that made it an abstract standard social contract (A-STD-SC). The story surrounding the other rule gave each proposition a clear meaning, but these meanings did not represent costs or benefits; hence, that problem was an ordinary abstract problem (AP). AV predicts that the percentage of 'P & not-Q' responses will be equally low on both problems; SC predicts a high percentage for the A-STD-SC and a low percentage for the AP.

Subjects.

Forty-two undergraduates from Harvard University participated in Experiment 5; they were volunteers from an introductory science course. Eighteen were in the group given the AP (13 females, 5 males), 24 in the group given the A-STD-SC (16 females, 8 males).

Materials and Procedure.

Although each test booklet contained five problems, only the abstract problems are relevant to Experiment 5. The fourth problem in every booklet was an abstract problem -- either the AP or the A-STD-SC. The other problems are shown in Figure 6.9. The first and third problems were food problems (one "salad bar" problem and one "Mexican food" problem per booklet), the second problem was a semi-social contract problem similar to Cox & Griggs' (1982) apparel color problem (see Chapter 2), and the fifth problem was a pilot threat problem (see Appendix C). Experiment 5 had a between-groups design.

Experiment 5 was part of a large battery of tests given at

Most of your friends find Mexican food too spicy. That's about all you knew about Mexican food when the Association of Mexican-American Restauranteurs (AMAR) first hired you to conduct market research on Americans' food preferences. So you started reading up on the subject. In the course of your reading, you came across a previously done report on the eating habits of Americans in Mexican restaurants which said:

"If a person eats hot chili peppers, then he will drink a cold beer."

You decide to look into this yourself. The cards below have information about four people in a Mexican restaurant in Boston. Each card represents one person. One side of a card tells what a person ate, and the other side of the card tells what that person drank.

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

A.	hot tea	B.	cold beer
C.	hot chili peppers	D.	broccoli

When the Consortium of Salad-Bar Owners & Operators (CSO&O) first hired you to conduct market research on Americans' food preferences, all you knew about salad bars was that your friends invariably start at the wrong end of them -- where the bacon bits & dressing are instead of where the lettuce is. So you started reading up on the subject. In the course of your reading, you came across a previously done report on the eating habits of Americans in salad bars which said:

"If a person eats lettuce, then he will drink water."

You decide to look into this yourself. The cards below have information about four people in a salad bar in Cambridge. Each card represents one person. One side of a card tells what a person ate, and the other side of the card tells what that person drank.

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

A.	water	B.	cheese
C.	coffee	D.	lettuce

Figure 6.9

← Food Problems

Apparel-Color Problem ↓

You are an anthropologist studying the Kaluame people, a Polynesian culture found only on Maku Island in the Pacific. The Kaluame adopted you into their culture and gave you the job of enforcing their societal laws. One of their social rules is:

"If a person is wearing a floral print shirt, then he must be over 20 years old."

The cards below have information about four Kaluame people sitting in a temporary camp; they don't seem to know one another. Each card represents one person. One side of a card tells what kind of a shirt a person is wearing and the other side of the card tells that person's age.

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

A.	solid black shirt	B.	floral print shirt
C.	42 years old shirt	D.	17 years old

the beginning of a class. Upon entering the class, each subject took a sealed test booklet from a stack containing booklets from the entire battery. The order of booklets in the stack was random. The first page of each booklet was an instruction sheet. The instructions were the same as for Experiments 1 through 4, except there was an additional sentence telling subjects that if a problem had more than one question associated with it, to answer the questions in the order they appear. Subjects had the opportunity to read the instructions to themselves; in addition, the instructions were read aloud to the entire class.

The two abstract problems are shown in Figure 6.10. Both rules express an alphanumeric code governing immigration documents (AP: "If a person has a 'D' rating, then his documents must be marked code '3'"; A-STD-SC: "If you fill out document 'D', then Ed will rate you code '3'"). To make sure subjects paid attention to what the letters and numbers stood for in the story, the selection task question was preceded by questions about each term. For the AP, these questions asked what each term meant; for the A-STD-SC they asked how the desires of characters in the story mapped onto each term.

The relation expressed in the AP was a sensible one -- if one translates the letters and numbers into the propositions they represent, it reads:
"If an applicant has submitted an incomplete medical record, then that applicant must see a health inspector."

I wanted both abstract rules to be set in a similar context -- thus both are bureaucratic rules involving immigration, and both are set in an immigration office. However, this constraint

Figure 6.10 Experiment 5: Abstract Immigration Problems

Abstract Standard Social Contract (A-STD-SC)

PAGE 4

You work for the national office of the U.S. Immigration Service, and it is your job to weed out corruption among immigration officers in the branch offices.

Four upstanding Mexican citizens want to immigrate to the U.S. Each needs to be rated code "3" by U.S. immigration in order to stay here; code "7" means instant deportation. Central Americans must file document D with immigration; Europeans must file document F. Each Mexican calls the Texas immigration office and talks to Betty, a very nice secretary who works for an immigration officer named Ed. Betty sends each a standard packet containing documents "A" through "H". And, in accordance with standard immigration law, Betty explains the law as follows:

"If you fill out document 'D', then Ed will rate you code '3'".

Ed is the very bigoted immigration officer who is in charge of processing these forms. You know that Ed hates Mexicans; he has made no secret of the fact that he does not want them coming into this country. You suspect that he will try to cheat some or all of these people out of their right to immigrate.

Questions:

1. Each Mexican wants to be rated code number 3.

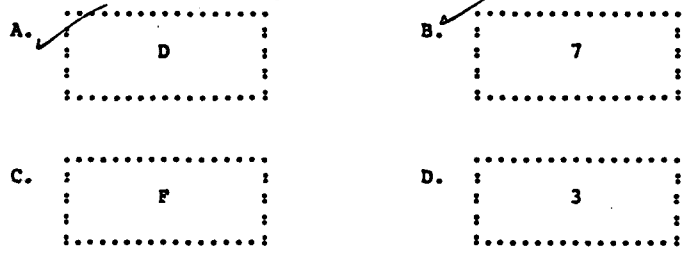
2. It is in the interest of each Mexican to fill out document D

3. The Mexicans would not need to fill out document F.

4. Personally, Ed would like to rate each Mexican code number 7.

5. The cards below have information about the documents of the four Mexicans who applied for immigration through Ed's office. Each card represents one person. One side of a card tells the letter name of the document the person filled out and the other side of the card tells what number code rating Ed gave that person.

The Washington office of the U.S. Immigration Service hired you to ferret out corruption in the branch offices. Ed is your target. Indicate only those card(s) you definitely need to turn over to see if Ed, by breaking the law Betty quoted, has cheated any of these Mexicans out of their right to immigrate.



irp

Abstract Problem

PAGE 4

Part of your new clerical job at the local immigration office is to make sure that the documents of people applying to immigrate into the U.S. are processed according to the proper alphanumeric rules relating letter ratings to number codes. There's a whole alphabet of letter ratings, and as many different number codes. For example, a "D" rating indicates that the applicant has submitted an incomplete medical record, an "F" rating means the applicant has dependent relatives; code "3" means the applicant must see a health inspector, code "7" means that the applicant must undergo a security check for possible criminal activities. The documents are already known to have the correct letter rating. Your job is to make sure the documents conform to the following alphanumeric rule:

"If a person has a 'D' rating, then his documents must be marked code '3'"

Questions:

1. Which code means the applicant must see the health inspector? 3

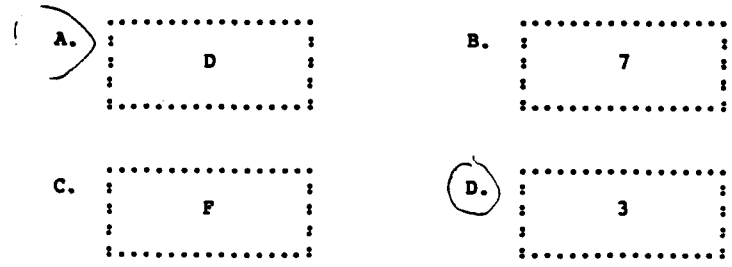
2. Which rating means the applicant's medical record is incomplete? D

3. Which rating means the applicant has dependent relatives? F

4. Which code means the applicant must undergo a security check? 7

5. You suspect the absent-minded secretary you replaced did not categorize the documents correctly. The cards below have information about the documents of four people who are applying for immigration. Each card represents one person. One side of a card tells a person's letter rating and the other side of the card tells that person's number code.

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



lq

made it difficult to construct a "full strength" social contract problem. By their nature, bureaucratic rules express "institutional desires", not the personal desires of the bureaucrat charged with enforcing them. What is a STD-SC from the institution's "point of view" may not be a proper social contract of any kind from the point of view of the bureaucrat administering it. This can be seen by translating the rule into cost/benefit terms from the point of view of the immigrants, the State, and Ed, the bureaucrat and potential cheater.

The opportunity to immigrate (be rated "code '3'") is considered a rationed benefit in this country; having to filling out papers is usually considered a cost/requirement. From the point of view of the immigrants' value system, the A-STD-SC rule translates to:

"If P then Q "
"If you pay the cost, then you are entitled to the benefit."

Ed, the potential cheater, is the bureaucrat charged with enforcing this rule. Cheating an immmigrant would clearly involve not giving him the benefit (not-Q) he is entitled to when he has paid the cost (P). Thus the correct SC answer is 'P & not-Q', making this problem a STD-SC. (Remember: a STD-SC only has the form "If B(X) then C(X)" when translated into the value system of the potential cheater. In this problem Ed is the potential cheater, not the immigrants.)

How does this rule translate from the point of view Ed, the potential cheater? A bureaucrat is charged with representing the desires of the institution that made the rule, not his own: the bureaucrat is supposed to behave as if his interests and the

interests of his employer coincide. This rule translates into a conventional STD-SC only if you view Ed as the instrument of the State -- an interpretation discouraged by the text. Subjects living in this country are presumably aware that the State -- Ed's employer -- views immigration as costly, but is willing to allow it when the person's character and finances are likely to provide benefits that offset the cost -- benefits assured, presumably, by filling out document D. If Ed, the potential cheater, is seen merely as the instrument of the State, then the rule is a proper STD-SC:

"If P then Q "
"If we receive the benefit, then we must pay the cost."

Translating the rule into Ed's personal value system is more difficult; because Ed did not make the rule, it does not necessarily express his values. He clearly agrees that the second term is a cost -- Ed doesn't want Mexicans entering the country. However, it is not clear that a completed "document 'D'" benefits Ed in any way. From Ed's point of view, the rule probably translates to:

"If P then Q "
"If O(Ed) then C(Ed)."

Moreover, we know from the story that Ed prefers not-Q (not letting Mexicans in) to Q (letting them in). Hence, in terms of cheating the immigrants, Ed has both motive and opportunity.

But note that he is not cheating them in the natural selection sense of the word -- he is not absconding with a benefit without paying the required cost. Ed is not benefited by

the immigrants having filled out document D*; by rating the immigrants code 7, Ed is avoiding the imposition of a cost, but not absconding with a benefit. From the immigrants' point of view, his doing this does cheat them in the natural selection sense of the word -- they have paid the cost, but not received the benefit that this entitles them to. Thus, in the natural selection sense, the immigrants have been cheated but Ed has not cheated them. This assymetry results from the fact that Ed did not himself make this contract with the Mexicans, so it does not express his own desires.

In fact, given Ed's bigotry, he probably does not see the rule as a social contract between himself and the immigrants, and thinks of it thus:

"If a Mexican fills out document 'F', then I get to rate him code '7'."
"If 0(Ed) then B(Ed) "

From the Mexican's point of view, this rule would be:

"If 0(Mexican) then 0(Mexican)"

-- not a social contract at all!

So, the rule is slightly confusing from a social contract point of view. If one identifies with the immigrants, then the conditions that constitute cheating are clear; if one identifies with Ed, they are not so clear. For this reason, performance may be weaker on this particular A-STD-SC than on the "full strength" U-STD-SCs used in Experiments 1 and 3.

---

* Nor is the State. Because B(State) resides in the character and finances of the potential immigrants, B(State) can only be collected if the Mexicans are allowed to immigrate. Thus, in the natural selection sense, the State, like Ed, cannot cheat the Mexicans -- it can, however, violate the rule.

Results.

Critical Test 7: Does an abstract standard social contract elicit the predicted SC response, 'P & not-Q'?

Percentage 'P & not-Q' responses:

	Social Contract Prediction:			Availability Prediction:		
A-STD-SC v. AP	A-STD-SC high	>	AP low	A-STD-SC low	=	AP low
	58%	>	22%			

The A-STD-SC elicited more "falsifying" responses than the AP, just as social contract theory predicts it should ($Z = 2.34$, $p < .01$, phi = .36). Hypothesis **AV** does not predict and cannot explain this result. However, the size of this effect is somewhat smaller than those found in Experiments 1 and 3 for the U-STD-SC v. AP comparisons -- about 57% the size (.36 for A-STD-SC v. AP, versus .61 and .65 for U-STD-SC v. AP in Experiments 1 and 3). There are at least three explanations for this smaller effect size:

1. Abstract symbols are more difficult to process than words, no matter how unfamiliar those words are. If this were true, then in Experiments 1-4, falsification rates would have been lower for AP problems than for U-D problems; they were not. Therefore, this explanation is rather unlikely.
2. Random variation in population sampling -- I have no way of assessing the merit of this explanation.
3. The A-STD-SC was not a "full-strength" social contract problem whereas the U-STD-SCs of Experiments 1 and 3 were (see discussion above).

I have no data that would allow me to choose between the second and third explanations.

Summary, Experiment 5.

An abstract standard social contract elicited a "content effect", as result predicted only by social contract theory. The effect was somewhat smaller than the social contract effects found in Experiments 1 through 4 for the unfamiliar problems.

Experiment 6

So far, unfamiliar and abstract social contract problems have all elicited high levels of predicted SC responses: these were the theoretically crucial problems for establishing a social contract effect and choosing between hypotheses **SC** and **AV**.

In Experiment 6, familiar STD-SC problems are pitted against abstract problems and familiar descriptive problems. Social contract theory predicts that a familiar STD-SC problem (F-STD-SC), like its unfamiliar analog, will elicit high levels of "logical falsification". Availability predicts middling performance on both F-D and F-STD-SC problems, unless, before seeing the results of the experiment, the individual availability theorist can concoct reasons why one problem or the other should prompt more counter-examples; so far, no principled way of making such judgments has been proposed (see Chapter 3).

The literature results reported in Chapter 2 indicate that familiar STD-SCs elicit replicable and robust social contract effects; however, given the elusivity of other content effects on the Wason selection task, it seemed prudent to test some F-STD-SCs on the same subject population that I used for the critical tests reported above.* Moreover, this allowed me to compare

* F-SWC-SCs cannot be used; experience with the more usual STD-SC form creates too many confounds (see Chapter 2: "Deformed Social Contracts").

performance on familiar STD-SCs to performance on unfamiliar ones, to see if familiarity produces an effect over and above the social contract effect (see above: "Social Contract Tests").

Experiment 6-A

The F-STD-SC used in this experiment was a Drinking Age Problem (DAP; see Chapter 2, Figure 4.1). The percentage of 'P & not-Q' responses elicited by the DAP was compared to that elicited by a transportation problem (F-D) and a prescriptive AP (shown in Figure 6.5).

Subjects.

Twenty-three undergraduates from Harvard University participated in Experiment 6-A (16 females, 7 males); they were volunteers from an introductory science course.

Materials and Procedure.

Experiment 6-A was part of the battery of tests described in Experiment 5. Each subject solved five problems, which appeared in the following order: AP, F-D, F-STD-SC (DAP), semi-SC (see Figure 6.9), and threat (see Appendix B). Only the first three problems are relevant to this analysis.

As discussed in the section on position effects, results indicate that falsification rates are not enhanced by increasing serial position; in fact, Cox & Griggs (1982) present evidence suggesting that first solving an AP incorrectly depresses falsification rates on a following DAP. If anything, then, having the F-STD-SC in the third position should narrow the difference between it and the AP. Moreover, having it follow the F-D eliminates the possibility of transfer from a correctly

solved F-STD-SC to the F-D.

Throughout Experiment 6, availability predictions are based on Griggs & Cox's (1982) memory-cueing version of availability theory. Predictions based on Pollard's (1982) differential availability would fare much worse, because, for very familiar STD-SCs like the DAP (Experiement 6-A) and the restaurant problem (Experiment 6-B), falsifying instances, though available, should be less available than confirming instances (see Chapter 3). Other brands of availability put forth are not sufficiently well-specified to make any strong predictions.

So as not to confuse issues, the social contract predictions for F-D problems will assume no effect of availability, as before; however, social contract theory is silent on whether availabilty exercises an independent effect on such problems.

Results.

The results are pictured in Table 6.12.

Table 6.12 Experiment 6-A: Percent of subjects choosing 'P & not-Q' for each problem (n=23)

AP:	30%
F-D:	57%
F-STD-SC:	78%

The tests that follow are not critical tests because the two hypotheses, SC and AV do not make radically different predictions about these problems (although the predictions of social contract theory are more tightly specified than those of availability theory). Rather, they are offered to show that the results are

consistent with social contract theory, and that social contract status produces reliable, replicable effects.

Does a familiar standard social contract elicit the predicted SC response, 'P & not-Q'?

Percentage 'P & not-Q' responses:

	Social Contract Prediction:			Availability Prediction:		
F-STD-SC v. AP	F-STD-SC high	>	AP low	F-STD-SC high-mid	>	AP low
	78%	>	30%	78%	>	30%

Experiment 6-A replicates the many experiments in the literature showing a robust content effect for the DAP. This F-STD-SC elicited a high percentage of SC responses -- 78%. Moreover, the proportion of 'P & not-Q' responses elicited by the F-STD-SC is significantly higher than that elicited by the AP (78% v. 30%: $F_{1,22} = 21.08$, $p << .001$, $r = .69$). This result is not predicted by Pollard's differential availability theory (despite his claims -- see Chapter 3), but it is consistent with Griggs & Cox's memory-cueing version of availability theory.

Are there more SC responses to a familiar standard social contract problem than falsifying responses to a familiar descriptive problem?

Percentage 'P & not-Q' responses:

	Social Contract Prediction:			Availability Prediction:		
F-STD-SC v. F-D	F-STD-SC high	>	F-D low	F-STD-SC high-mid	≥	F-D mid-low
	78%	>	57%	78%	>	57%

As social contract theory predicts, the F-STD-SC problem

elicited significantly more 'P & not-Q' responses than the F-D (78% v. 57%: $F_{1,22} = 6.11$, $p < .025$, $r = .47$). This result is also consistent with Griggs & Cox's availability theory, assuming that more subjects had experienced counter-instances to the DAP than to the transportation problem. It is not consistent with Pollard's differential availability.

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

	Effect:	No effect:
F-D v. AP	F-D > AP 57% > 30%	F-D = AP

The F-D transportation problem elicited significantly more falsifying responses than the AP (57% v. 30%: $F_{1,22} = 7.76$, $p <$

Experiments 1 through 4. As before, the availability effect is smaller than the social contract effect.

Experiment 6-B

Instead of the DAP, Experiment 6-B uses a F-STD-SC that involves discovering who has been cheating on the bill at a restaurant; other than that, it is identical to Experiment 6-A. The situation -- one person contributing less than her share when a group of friends eats out -- is probably familiar to college students.

This restaurant problem involves a simultaneous exchange: SC effects on the Wason selection task can be somewhat weaker for simultaneous, face-to-face exchanges, as they imply the possibility of intercontingent behavior that would prevent

cheating (see Chapter 5 and the discussion in Experiment 3). Unfortunately, I was not thinking about this when I created the problem. Nonetheless, it should elicit a social contract effect.

Subjects.

Twenty-four undergraduates from Harvard University participated in Experiment 6-B (17 females, 6 males, no data on sex of one subject); they were volunteers from an introductory science course.

Materials and Procedure.

The restaurant problem is shown in Figure 6.11. The other materials and the procedure were identical to those described for Experiment 6-A.

Results.

The results are pictured in Table 6.13.

Table 6.13 Experiment 6-A: Percent of subjects choosing 'P & not-Q' for each problem (n=23)

AP:	12%
F-D:	29%
F-STD-SC:	62%

Does a familiar standard social contract elicit the predicted SC response, 'P & not-Q'?

Percentage 'P & not-Q' responses:

	Social Contract Prediction:			Availability Prediction:		
F-STD-SC v. AP	F-STD-SC high	>	AP low	F-STD-SC mid-low?	>	AP low
	62%	>	12%			

Figure 6.11 Experiment 6-B: Restaurant Problem

Every Wednesday night you go out to dinner with the same group of four friends from work. You just got your first credit card, and you are trying to build up a good credit rating. So every time the five of you go out, you pay the check with your credit card, and they reimburse you on the spot with cash.

The problem is, the last few times you all have gone out, you've ended up paying more than your share of the bill. One of your friends has been consistently cheating you. This hurts your feelings, because you are the least well off of your friends -- you feel taken advantage of. So this Wednesday night you decide to figure out who is cheating you.

At the restaurant, some of your friends order hamburgers, and others splurge on lobster. When the check comes you announce the following rule:

"If you ordered lobster, then you must put in $20."

The cards below have information about your four friends. Each card represents one person. One side of a card tells what a person ordered and the other side of the card tells how much money that person kicked in.

Indicate only those card(s) you definitely need to turn over to see if any of your friends have broken the rule you announced.

A.

$10

B.

$20

C.

lobster

D.

hamburger

The F-STD-SC elicited the predicted social contract effect (62% v. 12%: $F_{1,23} = 23.00$, $p << .001$, $r = .71$). The problem's simultaneity does not appear to have diminished the size of the social contract effect -- .71 for the restaurant problem, .69 for the DAP.

It is difficult to say whether or not this result is also consistent with availability theory. Although the situation should qualify as familiar, so is eating while drinking, yet many theorists argue (post-hoc) that the food problem should not elicit a content effect (see Chapter 3). Whether AV predicts such a strong content effect for the restaurant problem depends on whether the particular availability theorist believes being shorted at a restaurant is a sufficiently common experience.

This result is important precisely because the content effect on the Wason selection task has been so elusive and unpredictable. The restaurant problem is a brand new problem, never tried before. It was designed according to the specifications of social contract theory and it elicited the predicted social contract effect. It is a new addition to the arsenal of successful social contract rules.

Are there more SC responses to a familiar standard social contract problem than falsifying responses to a familiar descriptive problem?

Percentage 'P & not-Q' responses:

	Social Contract Prediction:			Availability Prediction:		
F-STD-SC v. F-D	F-STD-SC high	>	F-D low	F-STD-SC mid-low?	=	F-D mid-low
	62%	>	29%			

According to social contract theory, the restaurant problem should elicit more 'P & not-Q' responses than the F-D transportation problem; this prediction was borne out (62% v. 29%: $F_{1,23} = 8.36$, $p < .01$, $r = .52$). Whether this result is also predicted by Griggs & Cox's memory-cueing depends on whether one predicts that more subjects have been shorted by a friend than have gone to Boston or Arlington by the form of transportation not mentioned in the rule -- I have no intuitions on this score. However, as it is likely that most subjects have had more experiences with their friends paying their fare share than not, this result contradicts Pollard's differential availability theory.

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

	Effect:	No effect:
F-D v. AP	F-D > AP 29% > 12%	F-D = AP

Availability appears to have produced a small but significant content effect for the F-D (29% v. 12%: $F_{1,23} = 4.60$, $p < .05$, $r = .41$). However, the social contract effect is larger, just as it has been in all previous experiments.

Taking Experiments 6-A and 6-B together and using the AP as a baseline, the size of the social contract effect is $r = .70$, and the size of the availability effect is $r = .46$. In other words, familiar social contract problems produce a social contract effect that is 1.52 times as big as the availability effect. This replicates the 50% advantage that social contract

status gave to unfamiliar problems in Experiments 1 through 4.

Experiments 6-A and 6-B versus Experiments 1-4: Familiar versus Unfamiliar Social Contract Problems

Social contract status has a profound effect on the percentage of SC answers elicited by a rule, but familiarity does not appear to further enhance this effect. The percent of SC responses to the unfamiliar SC problems in Experiments 1 through 4 does not differ significantly from those for the familiar SC problems of 6-A and 6-B; in fact, the unfamiliar problems elicited slightly more SC responses than the familiar ones (U-SC = 72%, F-SC = 66%: Z = 0.72, n.s.). Moreover, the effect sizes for the U-SC v. AP comparison are comparable to those for the F-SC v. AP comparison (U-SC v. AP, Experiments 1-4: .61, .69, .65,

sizes for the U-STD-SC v. F-D and F-STD-SC v. F-D comparisons are also comparable (U-STD-SC v. F-D, Experiments 1 and 3: .43, .47; F-STD-SC v. F-D, Experiments 6-A and 6-B: .47, .52). Either performance is so high for unfamiliar SC problems that there is a ceiling effect, or SC algorithms operate just as smoothly in unfamiliar situations as they do in familiar ones.

Experiment 6-C

The results so far have shown that the social contract effect is so large that the percentage of SC responses usually outstrips the percentage of falsifying responses to familiar descriptive problems. The transportation problem was the F-D problem used in all the previous experiments, because this was the non-SC problem that had been most successful in eliciting

content effects in the existing literature. To show that there is nothing special about the transportation problem, in this section performance on the DAP and restaurant problems -- the two F-STD-SC problems -- is compared to performance on an array of other familiar descriptive problems. The large battery of tests described in Experiment 5 allowed this comparison.

Subjects.

In addition to the 47 subjects who solved the DAP and restaurant problems of 6-A and 6-B, 115 Harvard undergraduates participated in Experiment 6-C (71 females, 42 males, no data on sex of two subjects); they were volunteers from an introductory science course.

Materials and Procedure.

All subjects solved five problems, but the only problems relevant to this analysis are those that fell in the third position, the same position as the F-STD-SC problems of 6-A and 6-B. The F-D problem for one group (n=23) was a rule relating a person's appearance in a particular kind of attire to his age; for another group (n=49) the F-D rule related arthritis to age -- these problems are shown in Figure 6.12. A third group (n=43) had one of the two food problems described in Experiment 5 as an F-D. The first two groups had an AP and transportation problem in the first two positions; in the third group, those positions were occupied by the other food problem and the semi-SC problem (Figure 6.9). Unlike the SC v. F-D comparisons of the previous experiments, this was a between-groups design, so the tests are not as powerful.

Figure 6.12 Experiment 6-C: Familiar Descriptive Problems

Arthritis Problem (F-D)

PAGE 2

You are a physician interested in disease in other cultures. You are studying the Kaluame people, a Polynesian culture found only on Maku Island in the Pacific. In medical school, you learned the following:

"If a person has arthritis, then he must be over 20 years old."

The cards below have information about four Kaluame people in a hospital. Each card represents one person. One side of a card tells what disease a person has and the other side of the card tells that person's age.

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

A. 15 years old

B. 58 years old

C. arthritis

D. tonsilitis

Appearance Problem (F-D)

PAGE 3

You are a fashion designer interested in the dress of other cultures. You are studying the dress habits of the Kaluame people, a Polynesian culture found only on Maku Island in the Pacific. In design school, you studied various sorts of fabrics and learned the following about floral prints:

"If a person looks fat in a floral print shirt, then he must be over 20 years old."

The cards below have information about four Kaluame people sitting around a campfire. All four are wearing floral print shirts. Each card represents one person. One side of a card tells how heavy a person looks and the other side of the card tells that person's age.

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

A. 15 years old

B. 35 years old

C. looks fat

D. looks thin

Results.

Table 6.14 displays the results. Both social contract problems elicited more 'P & not-Q' responses than each of the familiar descriptive problems. In four out of six tests this difference was significant. In two tests, those using the arthritis problem as an F-D, the difference is in the right direction but not significant. Note, however, that the percent of subjects who falsified on the arthritis problem is not outside the range found for the transportation problem with this subject population (arthritis: 57%; transportation: 58% (Exp 4), 50% (Exp 2), 57% (Exp 6-A)).

Table 6.14 Experiment 6-C: F-STD-SC v. F-D (non-transportation), Percent of subjects choosing 'P & not-Q'

			Drinking-Age Problem: 78% (n=23)
F-D	versus		F-STD-SC
Appearance:	41%	(n=49)	Z = 2.97, p < .0025, phi = .35
Food:	35%	(n=43)	Z = 3.97, p < .00005, phi = .49
Arthritis:	57%	(n=23)	Z = 1.57, n.s.
			Restaurant Problem: 62% (n=24)
F-D	versus		F-STD-SC
Appearance:	41%	(n=49)	Z = 1.74, p < .05, phi = .20
Food:	35%	(n=43)	Z = 2.18, p < .025, phi = .27
Arthritis:	57%	(n=23)	Z = 0.42, n.s.

Summary, Experiment 6.

Two familiar standard social contracts, including one that had never been tested before, elicited the predicted social contract effects. Not only did they elicit more "falsifying" responses than abstract problems, but they also elicited more than did a wide variety of familiar-descriptive problems.

Chapter 7

Discussion and Conclusions

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

Present in any novel situation are an infinite number of properties and relations. Darwinian algorithms are learning mechanisms keyed to focus attention on those dimensions of a situation that are evolutionarily important, and operate on them with inferential procedures that embody an appropriate evolutionary strategy. Without Darwinian algorithms, nothing could be learned; experience could not be structured to guide action along adaptive paths.

This thesis has proposed the existence of social contract algorithms. These focus attention on the actions of individuals, discern what those actions mean in terms of their desires, and calculate whether the cost/benefit structure of those desires indicates that the situation is one of social exchange; if it is, they operate on the cost/benefit structure of the situation with inference procedures that define cheating and facilitate the detection of cheaters. They operate in novel situations, as well as familiar ones, guiding inference and choices along adaptive pathways. The hypothesis that humans have social contact algorithms was tested using the Wason selection task, a test of how humans reason.

It was already known that how humans reason on the Wason selection task varies with what they are reasoning about; the question was, can the social contract hypothesis explain much of

that variation? The null hypothesis from the standpoint of the existing literature is that availability is the sole determinant of performance on Wason selection tasks of varying content. This was tested against the hypothesis that humans have social contract algorithms that are the major determinant of performance on Wason selection tasks whose content involves social exchange.*

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

For all six tests, the social contract hypothesis was verified and the null hypothesis that availability is the sole determinant of responses was falsified. Each of these six tests was replicated, using different unfamiliar social contract

* The social contract hypothesis is silent on whether availability exerts an indepedent effect on familiar problems that do not involve social exchange.

problems. The six critical tests, and subsequent experiments, established the following points:

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

The hypothesis that humans have social contract algorithms

uniquely accounts for the results of these experiments. It predicts them and it explains them; no other hypothesis proposed so far can do either. Moreover, the apparently contradictory literature attempting to stalk the "elusive" content effect on the Wason selection task can be systematically explained only by the social contract hypothesis.* Robust and replicable content effects are found only for rules that are standard social contracts: the only rules for which the predicted SC response is also the logically falsifying response.**

7.2 Are social contract algorithms innate?

Theoretical considerations

Availability theories presume the existence of innate learning mechanisms that are general purpose and content-independent. However, no variant of availability theory can adequately explain the results of the experiments presented in Chapter 6. It is difficult to see how the association 'cassava root-no tattoo' or 'eats duiker meat-has never found ostrich eggshell' could have been "cued" from long-term memory (Manktelow & Evans, 1979; Griggs & Cox, 1982), let alone be the dominant association for over 70% of undergraduates tested (Pollard, 1982). No matter how wildly unfamiliar the rule's terms, social contract problems elicited social contract responses. Furthermore, if associations between specific terms were responsible for the pattern of results on social contract rules,

---

* Problems with other proposed hypotheses are discussed in depth in Chapter 3.

** See Chapter 2 for a detailed review (especially p. 70-71).

then descriptive rules using the same unfamiliar terms should have elicited the same pattern; they did not. No theory whose predictive and explanatory power rests on associations between specific terms used in a social contract rule can explain the results of my experiments.

Availability theories that emphasize the role of mental modeling (Johnson-Laird, 1982) or frames (Wason, 1983; Rumelhart & Norman, 1981) in recognizing logical contradiction cannot explain the following aspects of the results:

1. Why would subjects find an unfamiliar social contract scenario (U-STD-SC) so much easier to model than an unfamiliar descriptive scenario (U-D), or indeed, a familiar one (F-D)?
2. Why would this situation reverse itself on switched social contracts (U-SWC-SC), for which the scenario to be modeled is identical to that for the U-STD-SC? Unlike the U-STD-SC, the U-SWC-SC does not elicit logically falsifying responses -- although it does elicit the correct social contract response.

The only response an availability theorist of the modeling variety could make would be to claim that people have a generalized social contract "frame" that recognizes and operates on the cost/benefit structure of a social contract as presented in Figure 6.1 (with which I agree -- see Chapter 5), but that it was acquired exclusively through "experience" -- more precisely, through experience structured solely by the content-independent, general purpose information processing systems presumed by associationists (Fodor, 1983). Innateness per se is not the

issue here: Every psychological theory -- even Hume's associationism -- assumes the existence of innate algorithms that structure experience. The issue is: Are some of the innate algorithms special purpose, content-dependent, Darwinian algorithms?

There is nothing in availability theory that would lead one to predict the existence of generalized social contract frames. Besides being post-hoc, this view cannot cope with a variety of fundamental issues, for example:

1. There is no reason to believe that SC rules are more common than non-SC prescriptive rules (bureaucratic rules, work orders from employers, safety rules, traffic rules, etc.) or the ubiquitous descriptive relations people use to describe and act on the world. Why, then, would general purpose algorithms have produced generalized SC frames, but not generalized frames for reliably detecting violations of descriptive or prescriptive* rules?

2. Compliance with social contracts is far more common than cheating. Every time a store lets you walk out with the goods you have paid for, you have experienced compliance. General purpose learning mechanisms should therefore create frames that look for compliance, not cheating.** At best, a

---

* A number of non-SC prescriptive rules have been tested, both in this thesis and in the literature: the AP immigration rule of Experiment 5; the AP of Experiments 1-4, 6; D'Andrade's AP; non-SC post-office rules in Golding, 1981 and Griggs & Cox, 1982; the apparel-color problem of Cox & Griggs, 1982; the deformed SC rules of Griggs & Cox, 1983 (see Chapter 2).

** Contextual exhortations to "look for cheaters" cannot explain this. Each non-SC problem contained several similar requests to see if the facts violate the rule, yet most subjects behaved like verificationists, nevertheless.

subject's ratio of compliance-to-cheating episodes should be the idiosyncratic product of different life experiences, and unfamiliar social contract problems should show at least as much response variability as familiar descriptive ones.

3. Trial and error learning requires some definition of error; hypothesis testing requires some definition of violation. A general purpose, content-independent learning mechanism needs a general purpose, content-independent definition of error. Logical falsification, for example, is a content-independent definition of error or violation. But the definition of violation for social contracts is quite specific: Cheating is defined as absconding with a benefit when you have not paid the required cost. It conforms to no known content-independent definition of error; it certainly does not map onto logical falsification, as a consideration of switched social contracts shows. Without built-in, domain specific knowledge defining what counts as cheating, how could one develop a generalized social contract frame?*

An evolutionarily-based social contract theory handles issues like these with ease. Social contract theory not only provides the most parsimonious explanation of the data, but the assumption that some innate algorithms are special purpose and content-dependent is also more parsimonious from the standpoint of evolutionary theory. Social exchange is a domain for which

* It is not sufficient to say that people learn what counts as cheating because they feel irritated when they have been cheated and therefore attend to the irritating stimuli; that presumes they already know what constitutes cheating. Having been cheated was the stimulus that triggered the irritation in the first place.

the evolutionarily-predicted computational theory is complex, and the fitness costs associated with "errors" are large. Even if it were possible for a domain general information processing strategy to construct social contract algorithms -- and it is by no means clear that it is possible -- it is not reasonable to expect that natural selection would leave learning in such a domain to a general purpose mechanism. Successfully conducted social exchange was such an important and recurrent feature of hominid evolution, that a reliable, efficient cognitive capacity specialized for reasoning about social exchange would quickly be selected for. A general purpose learning mechanism would either be supplanted or used only for learning in other domains.

Ontogeny

Evolutionary considerations can also guide research into the ontogeny of the social contract algorithms (see Cosmides, 1980). The brain is a metabolically expensive organ; expensive cognitive capacities should not mature until the organism needs them, so that metabolic energy can be devoted to other kinds of growth.* Social contract algorithms are not useful to a child until its welfare depends on individuals whose fitness interests are not identical to its own -- individuals to whom it must offer a benefit to get a benefit. Until weaning, the interests of mother and child are identical; benefiting the infant benefits the mother equally. Weaning marks the beginning of the end of this coincidence of interest. It is a period of intense parent-

* Spurts in brain growth appear to be correlated with spurts in cognitive development (Epstein, 1974a,b).

offspring conflict in both humans* and other primates (Trivers, 1976; Shostak, 1981), a period when the child wants more investment than the parent, who is ready to invest in a new offspring, is willing to give. After being weaned from the breast, the child is weaned from its mother's side; its welfare depends increasingly on the behavior of the less related individuals with whom it is left. At this point, the ability to cajole, threaten, exchange, and negotiate become crucial.

Thus, evolutionary considerations suggest that the learning mechanisms that underlie the ability to engage in social exchange should begin to mature slightly before the age usually associated with the onset of weaning during most of human evolution. World-wide, the average age of weaning is age two (Whiting, personal communication**), and this figure agrees well with life history estimates from the San, the hunter-gatherer group whose way of life is currently believed to most resemble that of Pleistocene hunter-gatherers.

A mechanism that guides learning about social exchange should include features that allow the child to a) model other people's values, both by noting their emotional reactions and attempting to manipulate their behavior, b) categorize values according to who has them, c) be aware of its own abilities, as these determine what the child is capable of offering to others,

---

* Most markedly in cultures where the only other food sources are difficult for infants to digest; for example, weaning among the San, who eat a high proportion of fibrous bush food as adults, appears to be particularly stressful (e.g., Shostak, 1981).

** This figure agrees with data for a population believed to approximate natural fertility conditions (Bongaarts & Potter, 1983, pp. 25, 90, 145).

d) understand and apply concepts of obligation and entitlement, e) become interested in notions of fairness and cheating, f) practice intercontingent behavior, g) remember its history of exchange with other individuals (see Chapter 5).

Intriguingly, Kagan (1981) has collected cross-cultural data suggesting the maturation, just prior to age two, of a cognitive capacity that looks suspiciously like it is specialized for learning about and engaging in social exchange. According to Kagan, the mental organ that emerges at this age includes:

* the concept of obligation,
* interest in and concern with other people's values,
* the ability to understand when an emotional reaction is "appropriate" to a person's age and situation,
* an awareness of one's own capacities for action,
* the ability to understand other people's intentions and anticipate their actions,
* an interest in trying to coax others into doing what the child wants (perhaps the most distinguishing characteristic of the "terrible twos").

Moreover, it is at about this age that language, the ultimate negotiative tool, begins to emerge. The computational theory of social exchange presented in Chapter 5 should allow one to generate predictions about other capacities that can be expected to co-occur with those already discovered (see Cosmides, 1980).

The finding that adult subjects are very adept at detecting potential "cheaters" on a social contract, even when it is unfamiliar and culturally alien, stands in marked contrast to the repeated finding that they are not skilled at detecting the

potential invalidity of descriptive rules, familiar or unfamiliar. The ontogeny of the algorithms that produce these results remains an open question. It is possible that they are, in some carefully delimited sense, learned. However, the mental processes involved appear to be powerfully structured for social contracts, yet weakly structured for other elements and relations drawn from common experience. This implies that the learning process involved is guided and structured by special purpose innate algorithms, just as learning a natural language is guided and structured by the innate algorithms of the language acquisition device.

7.3 The role of evolutionary theory in psychology

For the past 100 years, domain general psychological mechanisms have been the Holy Grail of experimental psychology. Paradigms rose and fell, mentalism gave way to behaviorism gave way to mentalism, but, undaunted, the quest for an equipotential psyche continued.

And psychology, after a century of research, is not yet an integrated science.

There may be a connection between these two facts. It may be that the processes that govern attention, perception, memory, categorization, reasoning, and learning simply are not equipotential. The Grail of legend could not be found because it did not exist.

The human mind did not evolve to attend "in general", to remember "in general," to learn "in general". It evolved to attend <u>to</u> predators, <u>to</u> the needs of kin, <u>to</u> potential sexual

partners, to agents of threat. The cognitive processes required for different evolutionarily important domains are different in kind: Attention to predators requires a high level of false positives to cues indicating felids and snakes; attention to the needs of kin requires selective orientation to emotion cues emitted by relatives, and the mobilization of reasoning and investigational processes that allow one to infer what it is that they need.

Attention, perception, categorization, learning, memory, decision making, and reasoning cannot be studied in isolation from motivation, emotion, behavior, and social psychology. To do so is to carve the study of psychology into artificial units that will not hang together. All these aspects of human psychology must be mobilized in different ways to solve different adaptive problems. As Chapter 5 illustrated, cognition, motivation, emotion, and behavior all must play specific and well-defined roles in solving the various adaptive problems associated with social exchange.

The search for domain specific Darwinian algorithms promises to integrate psychology, precisely because it focuses on adaptive problems. Cognitive psychologists can begin addressing issues closer to the heart of human nature. The study of emotion and motivation can be welcomed back into psychological theory, as systems for mobilizing the appropriate Darwinian algorithms when the situation "calls for" them. The exile of behavior from cognitive theory can end, because the presumed purpose of adaptive thought is to produce adaptive behavior.

Psychology and evolutionary biology are sister disciplines.

The goal of evolutionary theory is to define the adaptive problems that organisms must be able to solve. The goal of psychological theory is to discover the information processing mechanisms that have evolved to solve them. Alone, each is incomplete for the understanding of human nature. Together, they are powerful: as I hope the research presented in this thesis demonstrates, understanding what adaptive problems the human mind was designed to solve is a great aid to discovering how it works.

An evolutionary psychology would proceed adaptive problem by adaptive problem, domain by domain. Many adaptive problems have already been defined by evolutionary biologists. The real challenge for psychologists is to develop experimental methods that will allow the outlines of the psychological mechanisms that solve these problems to be traced. Happily, cognitive psychologists are in an excellent position to do this, having already invented an impressive array of concepts and experimental methods for tracking complex information processing systems. The experiments reported in Chapter 6 are a first attempt at such an approach: they used an experimental paradigm that had already been developed by cognitive psychologists.

The hypothesis that the human mind is a equipotential information processing system has been entertained for one hundred years. It is time for a change. The human mind is not a machine that fell out of the sky, of unknown purpose. The human mind was designed by natural selection to accomplish specific, well-defined adaptive functions. An equipotential psyche cannot accomplish these functions. A cognitive science that ignores this reality is a cognitive science that will fail.

Appendix A:

The Frame Problem and So-Called "Constraints" on Learning

Biologists and psychologists have a mysterious tendency to refer to the properties of domain specific (but not domain general) mechanisms as "constraints." For example, the one-trial learning mechanism, discovered by Garcia & Koelling (1966), that permits a blue jay to associate a food taste with vomiting several hours later, is frequently refered to as a "biological constraint on learning". Books reporting the existence of domain specific learning mechanisms frequently have titles like: "Biological Boundaries of Learning" (Seligman & Hager, 1972) or "The Tangled Wing: Biological Constraints on the Human Spirit" (Konner, 1982). This terminology is dangerously misleading, because it incorrectly implies that "unconstrained" learning mechanisms are a theoretical possibility.

All constraints are properties, but not all properties are constraints. Calling a property a "constraint" implies that the organism would have a wider range of abilities if the constraint were to be removed.

Are a bird's wings a "constraint on locomotion"? Birds can locomote by flying or hopping. Wings are a property of birds that enables them to locomote by flying, but wings are not a "constraint on locomotion." Wings expand the bird's capacity to locomote -- with wings, the bird can fly and hop. Removing a bird's wings reduces its capacity to locomote -- without wings, it can hop, but not fly. Wings cannot be a constraint, because removing them does not give the bird a wider range of locomoting abilities. If anything, wings should be called "enablers",

because they enable an additional form of locomotion. Having them actually expands the bird's capacity to locomote.*

A thick rubber band placed such that it pins a bird's wings to its body is a constraint on the bird's ability to locomote: With the rubber band the bird can only hop; without it the bird can both hop and fly.

Similarly, there is no evidence that the domain specific mechanisms that permit one trial learning of an association between a taste and vomiting are "constraints on learning." Removing the specific properties that allow the efficient learning of this particular association, would not expand the bird's capacity to learn, it would reduce it. Not only would the blue jay be unable to associate an electric shock with vomiting, it would also be unable to associate a food taste with vomiting.

Having wings to fly is, however, a constraint on (or more precisely, a restricted subset of) the theoretical class of all possible means of locomotion. A robin is capable of only two members of this theoretical set -- it cannot crawl, trot, roll, swim, burrow, or travel through time warps and worm holes in space -- it can only hop and fly. Having wings is not, however, a constraint on the organism's ability to locomote. Similarly, internal representations of the movements of solid objects appear to be "constrained" by the same laws of kinematic geometry that

---

* The ability to fly may turn out to place constraints on an alternative kind of locomotion, that is, efficient bipedal locomotion (flying requires hollow bones, which may not be strong enough to permit prolonged walking or hopping) but it is not a constraint on the birds capacity to locomote. Furthermore: Whether the ability to fly places constraints on the efficiency of bipedal locomotion is an empirical claim: One cannot simply assume, a priori, that having the ability to locomote by one means reduces the efficiency of another kind of locomotion.

govern the movement real objects: we only imagine a subset of the theoretically infinite number of possible paths by which an object can travel between two points (Shepard, 1984). This subset is the same subset true of real objects. However, domain specific knowledge like this expands the our capacity to accurately model the world, it does not reduce it.

This mysterious tendency is perhaps the result of the mistaken notion that a tabula rasa is possible, that learning is possible in the absence of a great deal of domain specific innate knowledge. If true, then a property that "prepares" an organism to associate vomiting with a taste may preclude it from associating an electric shock with that taste. However, if an organism had a domain general associative mechanism, there is no reason why that mechanism should not work to pair taste with electric shocks. One would have to hypothesize that the presence of food somehow shut off the domain general mechanism -- and this is an empirical claim that would have to be demonstrated.

Appendix B: Natural Selection in Action

How many generations will it take for indiscriminate altruists to go extinct?

(see Chapter 5, p. 138)

Imagine a population with n "altruists" (individuals who always cooperate, regardless of whether they are playing another altruist or a cheater) and n "cheaters" (individuals who never cooperate), where n is a very large number.* For simplicity's sake, assume each individual reproduces asexually, produces 2 offspring in the absence of any exchange, then dies. Each individual plays one Prisoner's Dilemma game per generation, and this game affects the number of offspring produced according to the payoff matrix in Figure 7.2 (+1 = one more offspring, for a total of 3; -1 = one less than, for a total of 1, and so on). Whether a particular individual plays with an altruist (A) or cheater (C) is random, and therefore proportional to the percentage of the population which each represents. P(A) = probability of playing with an altruist and P(C) = probability of playing with a cheater.

Expected reproductive value
for an individual altruist = [5 x P(A)] + [0 x P(C)]

Expected reproductive value
for an individual cheater = [7 x P(A)] + [2 x P(C)]

	Absolute numbers		Percent of Population	
	A	C	A	C
Parental generation	n	n	50%	50%
F1	2.5 n	4.5 n	36%	64%
F2	4.5 n	17.1 n	21%	79%
F3	4.7 n	52.2 n	8%	92%
F4	1.9 n	125.3 n	1.5%	98.5%
F5	.14 n	260.0 n	.05%	99.95%
F6	.00035 n	520.6 n	.000067%	99.999933%

When n = 10, the altruists are extinct after the fifth generation (F5).
When $10 < n < 2857$, the altruists would be extinct after the sixth generation (F6).

* This assumption simply smooths out the probabilities. For example, if n=10, then P(A plays with C) = .53, P(A plays with A) = .47, P(C plays with A) = .53, P(C plays with C) = .47. As n reaches infinity, all four probabilities converge on .5. Using the exact probabilities for a small n, simply drives altruists to extinction a bit faster.

Appendix C: Threat Problems

PAGE 5

You are a homicide detective for the Boston Police. For months, you have been gathering evidence against the infamous Owens Brothers. Jake and Ted Owens are in the drug trade, and are responsible for several particularly bloody underworld murders. They are shrewd and tricky -- they have eluded capture for months. You have amassed a huge amount of evidence against them -- your testimony in court could send them to jail for life. The problem is, they know it -- They have made several attempts on your life. They are ruthless killers and they want you dead.

An anonymous phone caller tells you that Ted will be down at the docks at 10 tonight. You go down there -- the docks are deserted. You turn a corner, and there is Ted. Quickly you pull your gun, shouting "Freeze!". Just as Ted is putting his hands in the air, you feel a gun in your back and hear Jake's cold voice behind you, saying:

"If you make one false move, I'll kill you."

What should you do? Does he mean it? The cards below have information about how this story could end. Each card represents an ending -- not necessarily different endings. One side of a card tells whether you gave up your gun, and the other side of the card tells whether Jake Owens shot you.

Indicate only those card(s) you definitely need to turn over to see if Jake has broken his "promise" in any of these endings.

A.	Jake shoots you	
B.	You give them your gun	
C.	Jake lets you go	
D.	You shoot Ted	

PAGE 5

You are a narcotics detective for the Cambridge Police. For months, you have been gathering evidence against Professor Owens and his student Bill. Professor Owens is a mild mannered fellow who is interested in the consciousness-expanding potential of hallucinatory drugs -- but he has had trouble getting his research funded. He and his students have been using University chemistry labs to manufacture LSD to sell on campus, in order to fund their research. You have enough evidence to arrest them -- the problem is, you have not been able to find them.

An anonymous phone caller tells you that the student, Bill, will be down at the docks at 10 tonight. You go down there -- the docks are deserted. You turn a corner, and there is Bill. Quickly you pull your gun, shouting "Freeze!". Just as Bill is putting his hands in the air, you feel a gun in your back and hear Professor Owens' voice behind you, saying:

"If you make one false move, I'll kill you."

What should you do? Does he mean it? The cards below have information about how this story could end. Each card represents an ending -- not necessarily different endings. One side of a card tells whether you gave up your gun, and the other side of the card tells whether Professor Owens shot you.

Indicate only those card(s) you definitely need to turn over to see if Professor Owens has broken his "promise" in any of these endings.

A.	Owens shoots you	
B.	You give them your gun	
C.	Owens lets you go	
D.	You shoot Bill	

BIBLIOGRAPHY

Personal communications:

David Pilbeam
Department of Anthropology
Harvard University

John Whiting
Department of Anthropology
Harvard University

Peter Wason
Department of Phonetics & Linguistics
University College London

E.O. Wilson
Department of Biology
Harvard University

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

Axelrod, R. & Hamilton, W.D. The evolution of cooperation. Science, 1981, 211, 1390-1396.

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

Bartlett, F.C. Remembering: A study in experimental and social psychology. Cambridge, UK: Cambridge University Press, 1932.

Boden, M. Artificial intelligence and natural man. New York: Basic Books, 1977.

Bongaarts, J. & Potter, R.G. Fertility, Biology, and Behavior. New York: Academic Press, 1983.

Bowlby, J. Attachment and loss (Volume 1). New York: Basic Books, 1969.

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

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

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

Bruner, J.S. In search of mind: Essays in autobiography. New York: Harper & Row, 1984.

Bruner, J.S., Goodnow, J.J. & Austin, G.A. A study of thinking. New York: Wiley, 1956.

Buss, D. Sex differences in human mate selection criteria: An evolutionary perspective. In Sociobiology and Psychology, in press.

Carey, S. & Diamond, R. Maturational determination of the developmental course of face encoding. In D. Caplan (ed.), Biological studies of mental processes. Cambridge, MA: The MIT Press, 1980.

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

Clark, H.H., & Clark, E.V. Psychology and language: An introduction to psycholinguistics. New York: Harcourt, 1977.

Cosmides, L. Adaptation and negotiation: A developmental approach. Unpublished manuscript, Harvard University, 1980.

Cosmides, L. Invariances in the acoustic expression of emotion during speech. Journal of Experimental Psychology: Human Perception and Performance, 1983, 9, 864-881.

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

Cutting, J.E., Proffitt, D.R., & Kozlowski, L.T. A biomechanical invariant for gait perception. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 357-372.

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

de Waal, F. Chimpanzee politics: Power and sex among apes. New York: Harper & Row, 1982.

Eibl-Eibesfeldt, I. Ethology: The biology of behavior (second edition). New York: Holt, Rinehart and Winston, Inc., 1975.

Ekman, P. Emotion in the human face (second edition). Cambridge, UK: Cambridge University Press, 1982.

Epstein, H.T. Phrenoblysis: Special brain and mind growth periods. I: Human brain and skull development. Developmental Psychobiology 7, 1974 (a), 207-216.

Epstein, H.T. Phrenoblysis: Special brain and mind growth periods. II: Human mental development. Developmental Psychobiology 7, 1974 (b), 217-224.

Evans, J.St.B.T., & Lynch, J.S. Matching bias in the selection task. British Journal of Psychology, 1973, 64, 391-397.

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

Fodor, J.A. The language of thought. Cambridge, MA: Harvard University Press, 1975.

Fodor, J.A. Modularity of mind. Cambridge, MA: The MIT Press, 1983.

Garcia, J. & Koelling, R.A. Relations of cue to consequence in avoidance learning. Psychonomic Science, 1966, 4, 123-124.

Gardner, H. The shattered mind. New York: Random House, 1974.

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

Gleitman, L.R. & Wanner, E. Language acquisition: The state of the state of the art. In E. Wanner & L.R. Gleitman (eds.), Language acquisition: The state of the art. Cambridge, UK: Cambridge University Press, 1982.

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

Goodall, J. van Lawick- The behaviour of free-living chimpanzees in the Gombe Stream Reserve. Animal Behaviour Monograph 3, 1968.

Goodall, J. van Lawick- In the shadow of Man. Boston: Houghton-Mifflin, 1971.

Goodwin, R.Q. & Wason, P.C. Degrees of insight. British Journal of Psychology, 1972, 63, 205-212.

Grice, H.P. Meaning. Philosophical Review, 1957, 66, 377-388.

Grice, H.P. Logic and conversation. Unpublished William James Lectures, Harvard University, 1967.

Griggs, R.A. The role of problem content in the selection task and in the THOG problem. In J. St.B.T. Evans, (Ed.), Thinking and reasoning: Psychological Approaches. London: Routledge & Kegan Paul, 1983.

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

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

Hall, K., & DeVore, I. Baboon social behavior. In I. DeVore (ed.), Primate behavior. New York: Holt, 1965.

Hamilton, W.D. The genetical evolution of social behaviour. Journal of Theoretical Biology, 1964, 7, 1-52.

Hamilton, W.D. Altruism and related phenomena, mainly in social insects. Annual Review of Ecology and Systematics, 1972, 3, 193-232.

Heisenberg, W. Physics and beyond. New York: Harper & Row, 1971.

Herrnstein, R.J. The evolution of behaviorism. American Psychologist, 1977, 32, 593-603.

Hughes, M.A.M. The use of negative information in concept attainment. (Doctoral dissertation, University of London, 1966.

Hume, D. An enquiry concerning human understanding. (E. Steinberg, ed.) Indianapolis: Hackett, 1977.

Hrdy, S. Blaffer. The woman that never evolved. Cambridge, MA: Harvard Univeristy Press, 1981.

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

Issac, G.Ll. The food-sharing behavior of protohuman hominids. Scientific American, 1978, 238, 90-108.

Janis, I., & Frick, F. The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. Journal of Experimental Psychology, 1943, 33, 73-77.

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

Johnson-Laird, P.N. Mental models: Towards a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press, 1983.

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

Johnson-Laird, P.N. & Wason, P. Insight into a logical relation. Quarterly Journal of Experimental Psychology, 1970, 22, 49-61.

Jolly, A. The evolution of primate behavior. New York: Macmillan, 1972.

Kagan, J. The emergence of self-awareness. Cambridge, MA: Harvard University Press, 1981.

Kant, I. Critique of pure reason. New York: Anchor Books, 1966.

Kinzey, W.G. Primate models for the origin of human behavior. New York: SUNY Press, 1985.

Konner, M. The tangled wing: Biological constraints on the human spirit. New York: Holt, Rinehart, & Winston, 1982.

Kozlowski, L.T., & Cutting, J.E. Recognizing the sex of a walker from a dynamic point-light display. Perception & Psychodynamics, 1977, 21, 575-580.

Landreth, B. Inside the inner circle. Popular Computing, 1985, 4, 62-65, 146-149.

Legrenzi, P. Relations between language and reasoning about deductive rules. In G.B. Flores D'Arcais & W.J.M. Levelt (Eds.), Advances in psycholinguistics. Amsterdam: North Holland, 1970.

Luce, R.D., & Raiffa, H. Games and decisions: Introduction and critical survey. New York: Wiley, 1957.

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

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

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

McCracken, G.F., & Bradbury, J.W. Social organization and kinship in the polygynous bat Phyllostomus hastatus. Behavioral Ecology & Sociobiology, 1981, 8, 11-34.

Minsky, M. Frame-system theory. R.C. Schank & B.L. Nash-Webber (Eds.), Theoretical issues in natural language processing, (unpublished, MIT), reprinted in P.N. Johnson-Laird & P.C. Wason (Eds.), Thinking: Readings in cognitive science. Cambridge, U.K.: Cambridge University Press, 1977.

Nozick, R. Philosophical explanations. Cambridge, MA: Harvard University Press, 1981.

Owens, J., Bower, G.H., & Black, J.B. The "soap opera" effect in story recall. Memory & Cognition, 1979, 7, 185-191.

Pollard, P. Human reasoning: logical and non-logical explanations. Unpublished PhD thesis, Plymouth Polytechnic, 1979.

Pollard, P. The effect of thematic content on the 'Wason selection task'. Current Psychological Research, 1981, 1, 21-29.

Pollard, P. Human reasoning: Some possible effects of availability. Cognition, 1982, 10, 65-96.

Pollard, P., & Evans, J. St B.T. The influence of logic on conditional reasoning performance. Quarterly Journal of Experimental Psychology, 1980, 32, 605-624.

Pollard, P. & Evans, J.St.B.T. The effects of prior beliefs in reasoning: An associational interpretation. British Journal of Psychology, 1981, 72, 73-81.

Popper, K.R. The logic of scientific discovery. London: Hutchinson, 1959.

Popper, K.R. Objective knowledge: An evolutionary approach. London: Oxford University Press, 1972.

Quine, W.V.O. Methods of logic. (Third Edition) New York: Holt, 1950.

Quine, W.V.O. Ontological relativity and other essays. New York: Columbia University Press, 1969.

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

Roberge, J.J. Linguistic and psychometric factors in propositional reasoning. Quarterly Journal of Experimental Psychology, 1978, 30, 705-716.

Roberge, J.J. Linguistic factors in conditional reasoning. Quarterly Journal of Experimental Psychology, 1982, 34, 275-284.

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

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

Schank, R. & Abelson, R.P. Scripts, plans, goals, and understanding. Hillsdale, NJ: Erlbaum, 1977.

Searle, J.R. (Ed.) The philosophy of language. Oxford: Oxford University Press, 1971.

Seligman, M.E.P. & Hager, J.L. Biological boundaries of learning. New York: Meredith, 1972.

Shepard, R.N. Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. Psychological Review, 1984, 91, 417-447.

Shepard, R.N. Evolution of a mesh between principles of the mind and regularities of the world. Paper presented at the Conference on Evolution and Information, Stanford University, April, 1985.

Shepher, J. Incest: A biosocial view. New York: Academic Press, 1983.

Shostak, M. Nisa: The life and words of a !Kung woman. Cambridge, MA: Harvard University Press, 1981.

Smuts, B. Special relationships between adult male and female olive baboons (Papio anubis). Doctoral dissertation, Stanford University, 1982.

Sokal, R.R., & Rohlf, F.J. Biometry. San Francisco: Freeman, 1969.

Strum, S.C. Baboons may be smarter than people. Animal Kingdom, 1985, 88, 12-25.

Tooby, J. Prospects for an evolutionary psychology. Unpublished manuscript, Harvard University, 1975.

Tooby, J. Pathogens, polymorphism and the evolution of sex. Journal of Theoretical Biology, 1982, 97, 557-576.

Tooby, J. The emergence of evolutionary psychology. In D. Pines (ed.), Emerging syntheses in science. New Mexico: Rio Grande Institute, in press.

Tooby, J. & DeVore, I. The reconstruction of hominid behavioral evolution through strategic modeling. In W.G. Kinzey (ed.), Primate models for the origin of human behavior. New York: SUNY Press, 1985.

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

Trivers, R.L. Parent-offspring conflict. American Zoologist, 1974, 14, 249-264.

Tversky, A. & Kahneman, D. Availability: A heuristic for judging frequency and probability. Cognitive psychology, 1973, 5, 207-232.

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

Van Duyne, P.C. Necessity and contingency in reasoning. Acta Psychologica, 1976, 40, 85-101.

Ward, P., & Zahavi, A. The importance of certain assemblages of birds as 'information-centres' for food finding. Ibis, 1973, 115, 517-534.

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

Wason, P.C. Reasoning about a rule. Quarterly Journal of Experimental Psychology, 1968, 20, 273-281.

Wason, P.C. Structural simplicity and psychological complexity: Some thoughts on a novel problem. Bulletin of the British Psychological Society, 1969, 22, 281-284. (a)

Wason, P.C. Regression in reasoning? British Journal of Psychology, 1969, 60, 471-480. (b)