The Ciba Foundation is an international scientific and educational charity (Registered Charity No. 313574). It was established in 1947 by the Swiss chemical and pharmaceutical company of CIBA Limited now Ciba-Geigy Limited. The Foundation operates independently in London under English trust law.

The Ciba Foundation exists to promote international cooperation in biological, medical and chemical research. It organizes about eight international multidisciplinary symposia each year on topics that seem ready for discussion by a small group of research workers. The papers and discussions are published in the Ciba Foundation symposium series. The Foundation also holds many shorter meetings (not published), organized by the Foundation itself or by outside scientific organizations. The staff always welcome suggestions for future meetings.

The Foundation's house at 41 Portland Place, London W1N 4BN, provides facilities for meetings of all kinds. Its Media Resource Service supplies information to journalists on all scientific and technological topics. The library, open five days a week to any graduate in science or medicine, also provides information on scientific meetings throughout the world and answers general enquiries on biomedical and chemical subjects. Scientists from any part of the world may stay in the house during working visits to London.

http://www.cibafoundation.demon.co.uk

Ciba Foundation Symposium 208

CHARACTERIZING HUMAN PSYCHOLOGICAL ADAPTATIONS

1997

JOHN WILEY & SONS

Chichester · New York · Weinheim · Brisbane · Toronto · Singapore

Dissecting the computational architecture of social inference mechanisms

Leda Cosmides and John Tooby

Center for Evolutionary Psychology, Department of Psychology, University of California, Santa Barbara, CA 93106, USA

Abstract. Scientists have been dissecting the neural architecture of the human mind for several centuries. Dissecting its computational architecture has proven more difficult, however. Within the cognitive sciences, for example, there is a debate about the extent to which human reasoning is generated by computational machinery that is domain specific and functionally specialized. While some claim that the same set of cognitive processes accounts for reasoning across all domains (e.g. Rips 1994, Johnson-Laird & Byrne 1991), others argue that reasoning is generated by several different mechanisms, each designed to operate over a different class of content (e.g. Baron-Cohen 1995, Cheng & Holyoak 1985, Cosmides & Tooby 1992, Leslie 1987). Indeed, it has recently been proposed that the human cognitive architecture contains a *faculty of social cognition*: a suite of integrated mechanisms, each of which is specialized for reasoning and making decisions about a different aspect of the social world. Candidate devices include a theory of mind mechanism, an eye direction detector, social contract algorithms, permission schemas, obligation schemas, precaution rules, threat detection procedures and others (e.g. Baron-Cohen 1995, Cheng & Holyoak 1985, Cosmides 1985, 1989, Cosmides & Tooby 1989, 1992, 1994, Fiddick et al 1995, Fiske 1991, Jackendoff 1992, Leslie 1987, K. Manktelow & D. Over, unpublished paper, 1st Int Conf on Thinking, Plymouth, UK 1988, Manktelow & Over 1990, M. Rutherford, J. Tooby, L. Cosmides, unpublished paper, 8th Annual Meeting Human Behav Evol Society, Northwestern Univ, IL 1996, J. Tooby & L. Cosmides, unpublished paper, 2nd Annual Meeting Human Behav Evol Society, Evanston, IL 1989). To decide among these sometimes competing proposals, psychologists need empirical methods and theoretical standards that let us carve social inference mechanisms at the joints. We will argue that the theoretical standards needed are those of the 'adaptationist programme' developed in evolutionary biology. To show how these standards can be applied in dissecting the computational architecture of the human mind, we will discuss some recent empirical methods and results.

1997 Characterizing human psychological adaptations. Wiley, Chichester (Ciba Foundation Symposium 208) p 132–161

The adaptationist stance

Focus on architecture

At a certain level of abstraction, every species has a universal, species-typical evolved architecture. For example, one can open any page of the medical textbook, *Gray's anatomy,* and find the design of this evolved architecture described down to the minutest detail — not only do we all have a heart, two lungs, a stomach, intestines and so on, but the book will also describe human anatomy down to the particulars of nerve connections. This is not to say there is no biochemical individuality: no two stomachs are exactly alike — they vary a bit in quantitative properties, such as size, shape and how much HCl they produce. But all humans have stomachs and they all have the same basic *functional* design — each is attached at one end to an oesophagus and at the other to the small intestine, each secretes the same chemicals necessary for digestion and so on. Presumably, the same is true of the brain and, hence, of the evolved architecture of our cognitive programmes — of the information-processing mechanisms that generate behaviour.

The cognitive architecture, like all aspects of the phenotype from molars to memory circuits, is the joint product of genes and environment. But the development of architecture is buffered against both genetic and environmental insults, such that it reliably develops across the (ancestrally) normal range of human environments. Characterizing the universal, species-typical architecture of human cognitive mechanisms is the central goal of psychology.

As psychologists, we are studying a system of fantastic complexity. Isolating a cog within an intricate machine of manifold interacting parts would be tremendously difficult, even if we had a large number of duplicates to experiment with. But we don't, because every time this fantastically complex system reproduces itself, sexual recombination injects variation into its design. Extracting what is invariant under these circumstances is a daunting task. It is not impossible, however. The structure of an evolved system reflects its function. Knowing a system's function can, therefore, illuminate its design. The 'adaptationist programme' is a research strategy in which theories of adaptive function are key inferential tools, used to identify and investigate the design of evolved systems.

Why does structure reflect function?

The evolutionary process has two components: chance and natural selection. Natural selection is the only component of the evolutionary process that can introduce complex *functional* organization into a species' phenotype (Dawkins 1986, Williams 1966).

The function of the brain is to generate behaviour that is sensitively contingent upon information from an organism's environment. It is, therefore, an information-processing device. Neuroscientists study the physical structure of such devices, and cognitive psychologists study the information-processing programmes realized by that structure. There is, however, another level of explanation — a functional level.

In evolved systems form follows function. The physical structure is there because it embodies a set of programmes; the programmes are there because they solved a particular problem in the past. This functional level of explanation is essential for understanding how natural selection designs organisms.

An organism's phenotypic structure can be thought of as a collection of 'design features'—micro-machines, such as the functional components of the eye or liver. Over evolutionary time, new design features are added or discarded from the species' design because of their consequences. A design feature will cause its own spread over generations if it has the consequence of solving adaptive problems: cross-generationally recurrent problems whose solution promotes reproduction, such as detecting predators or detoxifying poisons. If a more sensitive retina, which appeared in one or a few individuals by chance mutation, allows predators to be detected more quickly, individuals who have the more sensitive retina will produce offspring at a higher rate than those who lack it. By promoting the reproduction of its bearers, the more sensitive retina thereby *promotes its own spread over the generations* until it eventually replaces the earlier model retina and becomes a universal feature of that species' design.

Hence natural selection is a feedback process that 'chooses' among alternative designs on the basis of *how well they function.* It is a hill-climbing process, in which a design feature that solves an adaptive problem well can be outcompeted by a new design feature that solves it better. This process has produced exquisitely engineered biological machines—the vertebrate eye, photosynthetic pigments, efficient foraging algorithms, colour constancy systems—whose performance is unrivalled by any machine yet designed by humans.

By selecting designs on the basis of how well they solve adaptive problems, this process engineers a tight fit between the function of a device and its structure. To understand this causal relationship, biologists developed a theoretical vocabulary that distinguishes between structure and function.

Engineering standards

Those who study species from an adaptationist perspective adopt the stance of an engineer. In discussing sonar in bats, for example, Dawkins proceeds as follows: '... I shall begin by posing a problem that the living machine faces, then I shall consider possible solutions to the problem that a sensible engineer might consider; I shall finally come to the solution that nature has actually adopted' (Dawkins 1986, p 21–22).

Engineers figure out what problems they want to solve, and then design machines that are capable of solving these problems in an efficient manner. Evolutionary biologists figure out what adaptive problems a given species encountered during its evolutionary history, and then ask themselves, 'what would a machine capable of solving these problems well under ancestral conditions look like?' Against this background, they empirically explore the design features of the evolved machines that, taken together, comprise an organism. Definitions of adaptive problems do not, of course, uniquely specify the design of the mechanisms that solve them. Because there are often multiple ways of achieving any solution, empirical studies are needed to decide 'which nature has actually adopted'. But the more precisely one can define an adaptive information-processing problem—the 'goal' of processing—the more clearly one can see what a mechanism capable of producing that solution would have to look like. This research strategy has dominated the study of vision, for example, so that it is now commonplace to think of the visual system as a collection of functionally integrated computational devices, each specialized for solving a different problem in scene analysis—judging depth, detecting motion, analysing shape from shading and so on.

Design evidence

Because adaptations are problem-solving machines, they can be identified using the same standards of evidence that one would use to recognize a human-made machine: design evidence. One can identify a machine as a television rather than a stove by finding evidence of complex functional design: showing, for example, that it has many co-ordinated design features (antennas, cathode ray tubes, etc.) that transduce television waves and transform them into a colour bit map (a configuration that is unlikely to have arisen by chance alone), whereas it has virtually no design features that would make it good at cooking food. Complex functional design is the hallmark of adaptive machines as well. One can identify an aspect of the phenotype as an adaptation by showing that: (1) it has many design features that are complexly specialized for solving an adaptive problem; (2) these phenotypic properties are unlikely to have arisen by chance alone; and (3) they are not better explained as the by-product of mechanisms designed to solve some alternative adaptive problem. Finding that an architectural element solves an adaptive problem with 'reliability, efficiency and economy' is prima facie evidence that one has located an adaptation (Williams 1966).

Design evidence is important not only for explaining why a known mechanism exists, but also for discovering new mechanisms, ones that no one had thought to look for. Theories of adaptive function define what would count as a 'good design', and that allows one to generate testable hypotheses about the organization of a phenotypic structure. Thus, they can also be used heuristically, to guide investigations of phenotypic design.

Knowledge of adaptive function is necessary for carving nature at the joints

An organism's phenotype can be partitioned into adaptations, which are present because they were selected for; by-products, which are present because they are causally coupled to traits that were selected for (e.g. the whiteness of bone); and noise, which was injected by the stochastic components of evolution.

Lewontin (1979) defines adaptationism as 'that approach to evolutionary studies which assumes without further proof that all aspects of the morphology, physiology and behaviour of organisms are adaptive optimal solutions to problems'. By this definition, there are no adaptationists in the ranks of evolutionary biology. Not all aspects of an organism are functional, and every evolutionary biologist knows this (Williams 1966, Dawkins 1982).

Every machine, whether it was engineered by humans or by the evolutionary process, has non-functional aspects by virtue of being an ordinary causal system. The colour of the base of an overhead projector is unrelated to its function (projecting images on a screen) and so is the fact that the number of mirrors it has is a prime number. The colour is a by-product of a functional aspect (metals strong enough to support the projector happen to have a colour) and it has two mirrors because (given its function) the laws of reflectance require two — not because two is a prime number. The property, 'falls to earth when dropped', is a by-product of its having mass, and is most parsimoniously explained by appeal to the laws of gravitation.

Appeal to the laws of chemistry and physics are not sufficient, however, to explain why an overhead projector has mirrors, a light, a transparent surface and so on. These parts and properties are simultaneously present and arranged as they are *because* this configuration solves a problem. They are design features. To explain their presence and configuration, one needs to refer to the projector's function. Knowing its function is also necessary if one is to figure out which aspects of an overhead projector are *without* function. The same is true for organisms.

Like other machines, only narrowly defined aspects of organisms fit together into functional systems: most ways of describing the system will not capture its functional properties. Indeed, every organism has an infinite number of non-functional 'traits' because there are an infinite number of ways of carving a phenotype into 'parts' and 'properties' ('knee plus ear'; 'colour of mucus'; 'third epithelial layer of skin on the right arm plus salt receptors on tongue'; 'being less than 10 [or 20 or 2000 . . .] feet tall'; 'can be burned by acid'). For this reason, the assertion that organisms have non-functional aspects is true, but trivial.

Theories are developed to explain phenomena. The phenomenon that Darwin was trying to explain is the presence of *functional* organization in the phenotypes of organisms — the kind of organization that one finds in artefacts that were designed by an intelligent engineer to solve a problem of some kind. Functional organization is the *explanandum,* the phenomenon that the theory was developed to explain. Figuring out how to 'dissect' the architecture of a species in a way that illuminates this organization and explains its presence is, therefore, the one task that no Darwinian can escape or evade. To arrive at the appropriate construal, one must conceptualize this architecture as composed of non-random parts that interact in such a way that they solve adaptive problems. And this, of course, requires theories of adaptive function. They are engineering specifications, which provide the criteria necessary to decide whether a property of an organism is a design feature, a functionless by-product, a kluge in the system or noise.

Reverse engineering an inference system

The adaptationist programme can be used to reverse engineer inference systems. Its application suggests that the computational architecture of the human mind might be considerably different than is usually assumed (Cosmides & Tooby 1987, 1992, 1994, Tooby & Cosmides 1992).

Psychologists have long known that the human mind contains circuits that are specialized for different modes of perception, such as vision and hearing. But until recently, it was thought that perception and, perhaps, language were the only activities caused by cognitive processes that are functionally specialized. Other cognitive functions — learning, reasoning, decision making — were thought to be accomplished by circuits designed to operate uniformly over every class of content. These circuits were thought to be few in number, content independent and general purpose, part of a hypothetical faculty that generates solutions to all problems: 'general intelligence' (e.g. Fodor 1983, Johnson-Laird & Byrne 1991, Piaget 1950, Rips 1994). Experiments were designed to reveal what computational procedures these circuits embodied; prime candidates were all-purpose heuristics and 'rational' algorithms — ones that implement formal methods for inductive and deductive reasoning, such as Bayes's rule or the propositional calculus. These algorithms are jacks of all trades: because they are content free, they can operate on information from any domain (their strength). They are also masters of none: to be content independent means that they lack any domain-specialized information that would lead to correct inferences in one domain but would not apply to others (their weakness).

This research programme has produced a formidable paradox. When given artificial, laboratory-administered reasoning problems, people perform in ways that seem inept, especially when compared to artificial intelligence systems. Such findings led many psychologists to conclude that the faculty of human reasoning is riddled with crippling defects: heuristics, biases and fallacious principles that violate canons of rationality derived from logic, mathematics and philosophy (e.g. Kahneman et al 1982). Yet natural reasoning systems — human and non-human minds alike — negotiate the complex natural tasks of their world with a level of operational success far surpassing that of the most sophisticated existing artificial intelligence systems. Although artificial systems are usually composed of programmes that embody exactly those 'rational' principles that human minds are thought to lack, none has yet been able to match the performance even of a normal four-year-old child on everyday inferential tasks: inducing a grammar, analysing scenes, detecting predators, inferring the meaning of a smile, the wishes of a potential friend or the intentions of a potentially hostile stranger.

The paradox evaporates when one considers two things: (1) the limitations of rational algorithms; and (2) the nature of the problems human inference mechanisms were designed to solve.

Natural competences

An adaptationist would expect information-processing mechanisms — including inference systems — to be *ecologically rational*: to embody principles that allow adaptive problems to be solved with reliability, economy and precision (Tooby & Cosmides 1997). As a result, one expects them to work well under conditions that resemble the ancestral ones that shaped their design. They are calibrated to these environments, and they embody information about the stably recurring properties of these ancestral worlds.

One can think of the human computational architecture as a collection of evolved problem solvers. Many of these are expert systems, equipped with 'crib sheets': inference procedures and assumptions that embody knowledge specific to a given problem domain. These generate correct (or, at least, adaptive) inferences that would not be warranted on the basis of perceptual data alone. For example, there is now at least some evidence for the existence of inference systems that are specialized for reasoning about objects, physical causality, number, the biological world, the beliefs and motivations of other individuals, and social interactions (e.g. Atran 1990, Baron-Cohen 1995, Brown 1990, Cheng & Holyoak 1985, Cosmides 1989, Cosmides & Tooby 1989, 1992, Fiske 1991, Frith 1989, Hatano & Inagaki 1994, Jackendoff 1992, Keil 1994, Leslie 1987, 1988, Leslie & Thaiss 1992, Spelke 1990, Springer 1992, Wynn 1992).

Different problems require different crib sheets. For example, an assumption that is useful for predicting the behaviour of people — that their movements are caused by internal states, such as intentions, beliefs and desires — would be misleading if applied to inanimate objects. Two inference machines are better than one when the crib sheet that helps solve problems in one domain is misleading in another. This suggests that many evolved computational mechanisms will be domain specific: they will be activated in some domains but not others. Some of these may embody rational methods, but others will have special-purpose inference procedures that respond not to logical form but to content types — procedures that work well within the stable ecological structure of a particular domain, even though they might lead to false or contradictory inferences if they were activated outside of that domain.

An algorithm that is free of content is ignorant of the world. As a result, machines limited to executing Bayes's rule, *modus ponens,* and other procedures derived from mathematics or logic cannot go beyond the data of the senses. Having no crib sheets, there is little they can deduce about a domain; having no privileged hypotheses, there is little they can induce before their operation is hijacked by combinatorial explosion. The difference between domain-specific methods and domain-independent ones is akin to the difference between experts and novices: experts can solve problems faster and more efficiently than novices because they already know a lot about the problem domain.

Identifying domain-specific mechanisms

A major criterion for establishing the existence of a domain-specific inference system is whether, at an information-processing level, it appears to constitute a functionally

isolable computational unit. Is it activated independent of other units? Does it produce inferential steps unavailable to other units? Does it contain systems of procedures that are complexly specialized for processing information about a particular domain? Sometimes the neurological basis of a specialization can be identified and dissociated from other competences (see, for example, Leslie & Thaiss 1992, Baron-Cohen 1995, on dissociations between the theory of mind mechanism and other specializations), adding to the credibility of the cognitive-level characterization. But the primary criterion for distinguishing specializations is functional or cognitive, not neurological.

In reverse engineering a computational system composed of domain-specific inference engines, there are a number of questions that one must address.

(1) *Existence.* Does a hypothesized reasoning specialization actually exist, or are reasoning patterns better explained as the product of a domain-general mechanism in interaction with individual experiences and knowledge databases?
(2) *Scope.* What is the correct definition of the boundaries of the domain that the hypothesized reasoning unit operates over?
(3) *Proper cognitive description.* What is the correct specification of the specialization's procedures and representational formats?
(4) *Adaptive function.* Do these procedures and representational formats show the fit between form and function that one expects of a cognitive adaptation designed to solve the adaptive problem under consideration?
(5) *Universality.* Does it develop in all normal humans, regardless of culture, or is it sensitive to cultural variation and dependent on the details of individual experience?
(6) *Ontogenetic timing.* When does it develop — does it have a regular ontogenetic schedule?
(7) *Activation.* What conditions regulate its activation and deployment (e.g. can it be turned off and on by the presence or absence of a particular type of social situation)?
(8) *Regulation and function.* What activities are regulated or supported by the specialization? What other functions or behaviours are dependent on its output?
(9) *Inter-relationships.* What role does it play in a larger network of computational units? Do other specializations share a common database, or use its outputs as inputs or otherwise depend on its operation?
(10) *Neural basis.* Is the operation of the specialization (or its impairment) associated with particular regions of the brain? Is the specialization differentially activated or impaired by various hormones, drugs, or physiological and emotional states?
(11) *Role in real-world events.* What role does reasoning about a specific domain play in social interactions or other events? Does it explain aspects of real world phenomena (e.g. food sharing, gang-related violence)?
(12) *Health implications.* Does malfunctioning of the specialization (e.g. overactivation, underactivation, inappropriate activation) play a role in identifiable clinical disorders (e.g. autism, paranoia)?

As a result of research addressing these questions, there is growing evidence that the human cognitive architecture contains expert systems specialized for reasoning about the social world (e.g. Baron-Cohen 1995, Bugental & Goodnow 1997, Cheng & Holyoak 1985, Cosmides 1985, 1989, Cosmides & Tooby 1989, 1992, 1994, Etcoff et al 1991, Ekman 1992, Fernald 1992, Fiddick et al 1995, Fiske 1991, Jackendoff 1992, Leslie 1987, Mann 1992, M. Rutherford, J. Tooby, L. Cosmides, unpublished paper, 8th Annual Meeting Human Behav Evol Society, Northwestern Univ, IL 1996, J. Tooby & L. Cosmides, unpublished paper, 2nd Annual Meeting Human Behav Evol Society, Northwestern Univ, IL 1989). An adaptationist would expect their inference procedures, representational primitives and default assumptions to reflect the structure of adaptive problems that arose when our hominid ancestors interacted with one another. This expectation has guided our own research on human inference. We have proposed that reasoning about social exchange, precautions and threats is generated by three, functionally distinct, mechanisms; that each has a computational design that is specialized for solving the adaptive problems that typified its respective domain; and that each of these mechanisms is a component of the evolved architecture of the human mind — a reliably developing, species-typical set of cognitive procedures (e.g. Cosmides 1989, Cosmides & Tooby 1989, 1992, J. Tooby & L. Cosmides, unpublished paper, 2nd Annual Meeting Human Behav Evol Society, Evanston, IL 1989). We will use the literature on conditional reasoning to illustrate the role that the adaptationist programme can play in evaluating competing claims about the architecture of domain-specific reasoning mechanisms, and focus on social exchange as an example.

Characterizing the computational architecture that generates social inferences

Social computation and conditional reasoning

In categorizing social interactions, there are two basic consequences humans can have on each other: helping or hurting, bestowing benefits or inflicting costs. Some social behaviour is unconditional: for example, a mother nurses her infant without exacting a favour in return. However, most social acts are conditionally delivered. Indeed, much of the substance of human life is shaped by 'conditionals': statements or behaviours that express an intention to make one's behaviour contingent upon that of another. People conditionally help each other in reciprocal, dyadic, co-operative interactions; they conditionally threaten each other; and they form coalitions, defined by mutually understood contingencies of within-group co-operation and between-group competition. The inferential processes and decision rules that operate on conditionals make these activities possible and regulate their outcomes. This creates a selection pressure for cognitive designs that can detect and understand social conditionals reliably, precisely and economically (Cosmides 1985, 1989, Cosmides & Tooby 1989, 1992).

One important category of social conditional is social exchange — conditional helping — carried out by individuals or groups on individuals or groups. A social exchange involves a conditional of the approximate form: if person A provides the requested benefit to or meets the requirement of person or group B, then B will provide the rationed benefit to A (herein, a rule expressing this kind of agreement to co-operate will be referred to as a *social contract*).

Applying the adaptationist programme

Step 1: characterizing an adaptive problem. Our first step was to analyse the nature of such conditional interactions, including the structure of inferences that both: (1) make them possible; and (2) are necessary to guide an individual through these situations to successful outcomes or pay-offs. Using such analyses and the existing literature in economics, game theory and evolutionary biology, Cosmides (1985) and Cosmides & Tooby (1989) developed a task analysis or computational theory (in David Marr's sense) of the information-processing problems that arise in situations of social exchange. For example, economists and evolutionary biologists had already explored constraints on the emergence or evolution of social exchange using game theory, modelling it as a repeated Prisoners' Dilemma. One important conclusion was that social exchange cannot evolve in a species or be stably sustained in a social group unless the cognitive machinery of the participants allows a potential co-operator to detect cheaters (i.e. individuals who accept a benefit without satisfying the requirements that provision of that benefit was made contingent upon), so that they can be excluded from future interactions in which they would exploit co-operators (e.g. Axelrod 1984, Axelrod & Hamilton 1981, Boyd 1988, Trivers 1971, Williams 1966). Such analyses provided a principled basis for generating detailed hypotheses (called *social contract theory*) about reasoning procedures that would be capable, because of their domain-specialized nature, of detecting the presence of these social conditionals, interpreting their meaning and successfully solving the inference problems they pose. These hypotheses can be tested using standard methods from cognitive psychology.

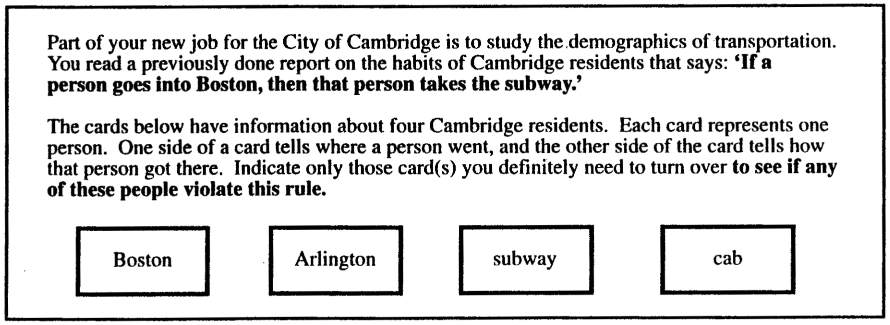
Step 2: searching for design evidence. Using the foregoing as a starting point, it was possible to engage the rich literature that already existed on how people reason about conditional rules — a literature that, in the early 1980s, lacked a single theory or set of theories that persuasively accounted for the known body of experimental findings. One of the principal tools reasoning researchers have used to explore conditional reasoning is the Wason selection task, a paper-and-pencil test in which subjects are asked to identify possible violations of a conditional rule of the form 'if P then Q'. Complex patterns in reasoning performance are elicited by differences in the content and context of conditional rules in the Wason selection task (e.g. Wason 1983, Wason & Johnson-Laird 1972). Such content effects are what one would predict if some reasoning was generated by domain-specific reasoning specializations, such that different procedures are activated by different contents.

Content-dependent performance on the Wason selection task. The Wason selection task was originally designed to see whether people are intuitive Popperians: whether they spontaneously attempt to falsify conditional rules by applying content-independent rules of logic. It is a word problem in which subjects are asked what additional information they would need to see to determine whether a conditional rule of the form 'if P then Q' has been violated by any one of four instances. Each instance is represented by a card. One side of a card tells whether the antecedent is true or false (i.e. whether P or not-P is the case), and the other side of that card tells whether the consequent is true or false (i.e. whether Q or not-Q is the case). The subject, who is only allowed to see one side of each card, is asked which card(s) must be turned over to see if any of them violate the rule. The four cards the subject must choose from show terms representing the logical categories P, not-P, Q and not-Q (Fig. 1). The rules of logical inference are content free so, no matter what P and Q stand for, the logically correct response is to choose the P card (to see if it has a not-Q on the other side) and the not-Q card (to see if it has a P on the other side).

There is a large body of literature showing that people are not good at detecting potential violations of conditional rules, even when these rules deal with *familiar content drawn from everyday life*. For example, descriptive rules—conditionals describing some state of the world—typically elicit a fully correct response (P and not-Q) from only 5–25% of subjects tested (Cosmides 1985, Wason 1983).

The Wason selection task is a convenient tool for testing hypotheses about reasoning specializations designed to operate on social conditionals because: (1) it tests reasoning about conditional rules; (2) the task structure remains constant while the content of the rule is changed; (3) content effects are easily elicited; and (4) there is already a body of existing experimental results against which performance on new content domains can be compared. For example, to show that people who ordinarily cannot detect violations of conditional rules can do so when that violation represents cheating on a social contract would constitute initial support for the view that people

have reasoning procedures specialized for detecting cheaters in situations of social exchange. To find that violations of conditional rules are spontaneously detected when they represent bluffing on a threat—a computationally different problem—would, for similar reasons, support the view that people have reasoning procedures specialized for analysing threats. Our general research plan has been to use subjects' inability to spontaneously detect violations of conditionals expressing a wide variety of contents as a comparative baseline against which to detect the presence of performance-boosting reasoning specializations. By seeing what content manipulations switch on or off high performance, the boundaries of the domains within which reasoning specializations successfully operate can be mapped. For example, there are now a number of experiments comparing performance on Wason selection tasks in which the conditional rule either did or did not express a social contract: a situation in which one is entitled to a benefit from a party only if one has satisfied the requirement that the offer of this benefit was made contingent upon (e.g. 'If a man eats cassava root [described as an aphrodisiac], then he must have a tattoo on his face'). Although very few subjects correctly identify potential violations of descriptive conditionals, 65–80% of subjects do so when the conditional rule expresses a social contract and a violation represents cheating. Subjects routinely check for cheating by choosing the cards that represent a person who has accepted the benefit ('ate cassava root') and a person who has not satisfied the requirement ('has no tattoo'). Furthermore, it is not just a question of how much 'facilitation' a conditional rule elicits: different hypothesized reasoning specializations predict different choices on the Wason selection task. The *pattern* of choices subjects make, given the content of the problem, can be used to test alternative hypotheses about the nature of the reasoning procedures activated. For example, the inference mechanisms that generate responses to social contracts *do not apply content-free logical rules*: they cause subjects to choose the benefit accepted card and the requirement not satisfied card regardless of their logical category (Fig. 2).

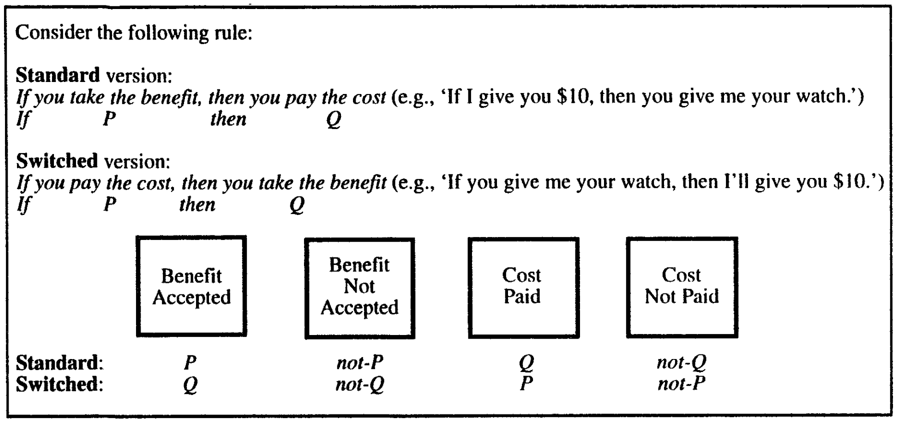

Part of your new job for the City of Cambridge is to study the demographics of transportation. You read a previously done report on the habits of Cambridge residents that says: **'If a person goes into Boston, then that person takes the subway.'**

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.**

Boston Arlington subway cab

FIG. 1. The Wason selection task (descriptive rule, familiar content).

In fact, experiments that systematically manipulate problem content demonstrate a series of domain-specific effects predicted by our computational theory of social exchange, thereby providing a substantial body of design evidence. They show that the mechanisms activated by social contract content have many components that appear to be functionally specialized for reasoning about social exchange, including procedures that are well designed for detecting cheaters. For instance: (1) these procedures operate so as to detect cheaters, even when the social contract expressed is highly unfamiliar; (2) they do not operate unless the representation of the rule satisfies the cost–benefit constraints of a social contract; (3) the only violations they detect are ones that represent illicitly taken benefits (cheating)—they are not good at detecting innocent mistakes; (4) they are sensitive to whose perspective is being taken in an exchange; (5) they do not embody a content-independent formal logic—they identify cheaters even when this leads to answers that violate the strictures of, for example, the propositional calculus; (6) they cause people to 'read in' deontic operators such as 'may' and 'must', corresponding to obligation and entitlement,



Consider the following rule:

Standard version:
If you take the benefit, then you pay the cost (e.g., 'If I give you $10, then you give me your watch.')
If P then Q

Switched version:
If you pay the cost, then you take the benefit (e.g., 'If you give me your watch, then I'll give you $10.')
If P then Q

	Benefit Accepted	Benefit Not Accepted	Cost Paid	Cost Not Paid
Standard:	*P*	*not-P*	*Q*	*not-Q*
Switched:	*Q*	*not-Q*	*P*	*not-P*

FIG. 2. Generic structure of a social contract.

even when these are not explicitly stated in the rule; (7) they can be primed separately from other kinds of deontic rules; and (8) the ability to correctly identify potential cheaters can be intact in individuals suffering from schizophrenia, even though schizophrenia causes impairments in logical and other deliberative reasoning tasks — a result that indicates that social exchange procedures are neurologically dissociable from mechanisms that govern 'general intelligence'. (For review, see Cosmides & Tooby 1992, 1989, Cosmides 1985, 1989, Fiddick et al 1995, Gigerenzer & Hug 1992, Maljković 1987, Platt & Griggs 1993). Wason selection tasks involving social exchange elicit a pattern of results so distinctive that we have proposed that reasoning in this domain is governed by computational units that are domain specific and functionally distinct: *social contract algorithms* (Cosmides 1985, 1989, Cosmides & Tooby 1992).

Step 3: eliminating by-product hypotheses. The human cognitive phenotype has many features that appear to be complexly specialized for solving the adaptive problems that arise in social exchange. However, demonstrating this is not sufficient for claiming that these features are cognitive adaptations *for* social exchange. One also needs to show that these features are not more parsimoniously explained as the by-product of mechanisms designed to solve some other adaptive problem or class of problems.

For example, Cheng & Holyoak (1985, 1989) also invoke content-dependent computational mechanisms to explain reasoning performance that varies across domains. But they attribute performance on social contract rules to the operation of a permission schema (and/or an obligation schema; these do not lead to different predictions on the kinds of rules usually tested; see Cosmides 1989), which operates over a larger class of problems. They propose that this schema consists of four production rules:

(1) if the action is to be taken, then the precondition must be satisfied;
(2) if the action is not to be taken, then the precondition need not be satisfied;
(3) if the precondition is satisfied, then the action may be taken; and
(4) if the precondition is not satisfied, then the action must not be taken;

and that their scope is any permission rule, that is, any conditional rule to which the subject assigns the following abstract representation: 'if action A is to be taken, then precondition P must be satisfied'. All social contracts are permission rules, but not all permission rules are social contracts. The conceptual primitives of a permission schema have a larger scope than those of social contract algorithms. For example, 'a benefit taken' is a kind of 'action taken', and a 'cost paid' (i.e. a benefit offered in exchange) is a kind of 'precondition satisfied'. They take evidence that people are good at detecting violations of precaution rules — rules of the form, 'if hazardous action H is taken, then precaution P must be met' — as evidence for their hypothesis (on precautions, see K. Manktelow & D. Over, unpublished paper, 1st Int Conf on Thinking, Plymouth, UK 1988, Manktelow & Over 1990). After all, a precaution rule is a kind of permission rule, but it is not a kind of social contract. We, however, have hypothesized that reasoning about precaution rules is governed by a functionally specialized inference system that differs from social contract algorithms and operates independently of them (Cosmides & Tooby 1992, Fiddick et al 1995).

In other words, there are two competing proposals for how the computational architecture that causes reasoning in these domains should be dissected. Application of the adaptationist programme suggests at least five kinds of evidence that can be used to decide between them.

(1) Over which transformations is the behaviour of the system invariant?

Transformations of input variables. Native speakers of English recognize that: (a) 'furry brown bears sleep soundly'; and (b) 'colourless green ideas sleep furiously' are both grammatical sentences with identical syntactic structures. Transformations that substitute alternative words from the same part of speech are irrelevant to this judgement. Transformations of phrase structure are not: although the words of 'furry soundly bears brown sleep'are identical to those in (a), it is not a grammatical sentence. With examples like these, Chomsky (1957) showed that English syntax uses arguments such as *noun* and *verb*, which must bear a certain relationship to one another. Moreover, the grammatical system must be independent of meaning systems, because its operation is invariant over transformations that change word meaning, but it must preserve syntactic structure.

Transformations of input should be irrelevant to the operation of *any* syntactic system, as long as they fall within the range of input variables that its arguments accept: i.e. as long as they do not violate its *argument structure*. By systematically varying input, one should be able to discover the rules of a syntactic system and the arguments they take. This principle should allow one to discover which proposed syntax — that of the permission schema or the social contract algorithms — better describes the behaviour of the inference systems activated by the problems discussed above.

For example, according to the grammar of social exchange, a rule is not a social contract unless it contains a *benefit to be taken*. Transformations of input should not matter, as long as the subject continues to represent an action or state of affairs as beneficial to the potential violator, and the violator as illicitly obtaining this benefit. This is true: performance on social contract rules is just as good when the benefit to be taken is highly unfamiliar (e.g. eating cassava root, getting an ostrich eggshell) as when it is familiar (e.g. drinking beer, being assigned to a good high school).

The corresponding argument of the permission schema — *an action to be taken* — has a larger scope: not all 'actions taken' are 'benefits taken'. If this construal of the rule's argument structure is correct, then the behaviour of the reasoning system should be invariant over transformations of input that preserve it. But it is not. For example, consider two rules: (a) 'if one goes out at night, then one must tie a small piece of red volcanic rock around one's ankle'; and (b) 'if one takes out the garbage at night, then one must tie a small piece of red volcanic rock around one's ankle'. Most undergraduate subjects perceive the action to be taken in (a) — going out at night — as a benefit, and 80% of them answered correctly. But when one substitutes a different action — taking out the garbage — into the same place in the argument structure, then performance drops to 44%. This transformation of input preserves the *action to be taken* argument structure, but it does not preserve the *benefit to be taken* argument structure — most people think of taking out the garbage as a chore, not a benefit. If the syntax of the permission schema were correct, then performance should be invariant over this transformation. But a drop in performance is expected if the syntax of the social contract algorithms is correct.

We have been doing similar experiments with precaution rules (e.g. 'if you make poison darts, then you must wear rubber gloves'). All precaution rules are permission rules (but not all permission rules are precaution rules). We have been finding that the degree of hazard does not affect performance, but the nature of the precaution does — even though all of the *precautions taken* are instances of *preconditions satisfied*. Performance drops when the precaution is not perceived as a good safeguard given the hazard specified (M. Rutherford, J. Tooby & L. Cosmides, unpublished paper, 8th Annual Meeting Human Behav Evol Society, Northwestern Univ, IL 1996, J. Tooby & L. Cosmides, unpublished paper, 2nd Annual Meeting Human Behav Evol Society, Evanston, IL 1989). This is what one would expect if the syntax of the rules governing reasoning in this domain take arguments such as *facing a hazard* and *precaution taken*; it is not what one would expect if the arguments were *action taken* and *precondition satisfied*.

Transformations of context. A syntactic system has certain operators and conceptual primitives. For example, in both the permission schema and social contract algorithms *must* is a deontic operator indicating obligation (not a modal indicating necessity). But social contract algorithms contain certain conceptual primitives that the permission schema lacks. For example, *cheating* is taking a benefit that one is not entitled to; we have proposed that social contract algorithms have procedures that are specialized for detecting *cheaters*. This conceptual primitive plays no role in the operation of the permission schema. For this schema, whenever the action has been taken but the precondition has not been satisfied, a *violation* has occurred. People should be good at detecting violations, whether that violation counts as cheating (the benefit has been illicitly taken by the violator) or a mistake (the violator does not get the benefit stipulated in the rule).

Given the same social contract rule, one can manipulate contextual factors to change the nature of the violation from cheating to a mistake. When we did this, performance changed radically, from 68% correct in the cheating condition to 27% correct in the mistake condition. Gigerenzer & Hug (1992) found the same drop in response to a similar context manipulation. (Their interesting 'perspective change' experiments provide another example of how transformations of context can be used to decide between competing proposals.)

(2) Neurological dissociations

Cognitive neuroscientists have been using data about double dissociations to dissect the cognitive architecture. Strokes, head traumas, diseases and developmental disorders (such as autism) sometimes damage one mechanism without affecting another. For example, as a result of brain damage, some people lose the ability to recognize individual faces, but they can still recognize emotional expressions on faces. Others lose the ability to recognize emotional expressions, but they can still discriminate individual faces (Etcoff 1984). This is called a *double dissociation*. It is prima facie evidence that performance on these two tasks is caused by two different mechanisms, rather than one. If it had been caused by one mechanism, then damaging that mechanism should depress performance on both tasks: it should not be possible to find people in which performance on each is *selectively* impaired.

The same logic can be applied to the study of reasoning. If reasoning about social contracts and precautions is caused by one and the same mechanism — a permission schema — then neurological damage to this schema should lower performance on both rules equally. But if reasoning about these two domains is caused by two, functionally distinct mechanisms, then one could imagine neurological damage to the social contract algorithms that leaves the precaution mechanisms unimpaired and vice versa. Selective damage of social contract algorithms should depress performance on social contracts, but not on precaution rules. Selective damage of precaution mechanisms should depress performance on precaution rules, but not on social contracts.

Strokes, head traumas, diseases and developmental disorders often produce extensive brain damage. So even if a two-mechanism hypothesis is correct, there is no guarantee that one will find a patient in which the damage is localized enough to impair one mechanism but not the other. But if such patients *can* be found, they provide prima facie evidence that reasoning in these two domains is caused by two separate mechanisms. In collaboration with Valerie Stone, we have been giving Wason selection tasks to individuals with focal brain damage. Although this effort has just begun, it would appear that we have found a patient who performs at high levels on precaution rules, but not on social contract rules (V. Stone, L. Cosmides & J. Tooby, unpublished paper, 8th Annual Meeting Human Behav Evol Society, Northwestern Univ, IL 1996). We will continue to screen people, to see if people with the opposite dissociation — performing at high levels on social contracts, but not on precaution rules — can also be found.

Complex machines can be 'broken' in a number of different ways, and the logic of double dissociations can be applied to impairments caused by genetic variation as well. Natural selection tends to eliminate genetic variation, so the heritability of adaptations is usually low, not high (for discussion of the relationship between heritability and adaptationism see Tooby & Cosmides 1990). Nevertheless, sexual recombination injects 'noise' into phenotypic development. As a result, 'normal' genetic variation has many random effects — sometimes ones that impair the functioning of one mechanism while leaving another intact (e.g. most people who are genetically predisposed to myopia speak normally). Such effects can assist in the dissection of computational architecture. For example, a mutation involving a single dominant autosomal gene impairs the acquisition of certain grammatical rules, yet has no effect on spatial reasoning and other forms of non-verbal intelligence (Gopnik 1990). Individuals with a different genetic disorder, Williams syndrome, speak fluently and grammatically, yet are severely retarded (Pinker 1994). This double dissociation indicates that the mechanisms responsible for non-verbal 'intelligence' are not sufficient to explain the acquisition of grammar, and vice versa: more than one mechanism must be invoked to explain performance in these two domains.

Genetic variation could, in principle, selectively impair one reasoning mechanism while sparing another, making behaviour genetic data relevant. It is informative to know, for example, whether the concordance of a particular reasoning dissociation is higher in identical than fraternal twins. An advantage of this method is that it vastly expands the pool of potential subjects, to encompass many people who would not ordinarily be classified as having neurological damage.

(3) Functional dissociations: evidence from priming

The logic of double dissociations can be extended to encompass *functional* (rather than neurological) dissociations. If one can create experimental conditions that temporarily activate (or deactivate) mechanisms in a selective way, one can see whether levels of performance on two different 'target' tasks can be dissociated. One way of doing this is through *priming*.

Priming experiments have been used extensively in memory research, sometimes to decide whether performance on two different tasks was caused by a single memory system, or two different ones. The subject engages in an activity that temporarily activates ('primes') a mechanism or representation; one then sees whether this influences performance on a target task that immediately follows.

In a typical priming experiment, the subject is given two tasks in sequence. The first is called the 'prime', and the second is called the 'target'. To see whether the initial task influences performance on the target task, one compares performance on the target when it follows the prime to performance on the target when it follows a control task. When performing the initial task enhances performance on the target task (by some measure, such as reaction time or percent correct), this is called 'priming': the initial task primed performance on the target task. Using the Wason selection task, we have adapted this method to the study of reasoning, and found that we can: (1) selectively activate functionally distinct inference mechanisms; and (2) thereby elicit reasoning by analogy.

If domain-specific inference engines structure how cognitive architectures construe similarity across situations, then one social situation will be categorized as the same as another if both can be mapped onto the same set of domain-specific representations. For example, an ambiguous rule such as 'if a person wears a grey shirt, then that person is 19 years old' will be categorized as a social contract if the context makes it clear that the people under discussion: (i) think of wearing grey shirts as a rationed benefit (e.g. a military honour); and (ii) construe the statement about age as a requirement. Otherwise, the subject will categorize the rule as one which simply describes the habits of people over 19, and it will elicit the low levels of performance generally found for descriptive rules. If context causes an otherwise ambiguous rule to be categorized as a social contract, then cheater-detection procedures can operate on it, generating high levels of performance. Previous experiments on social contract reasoning have confirmed this prediction (see, for example, Cosmides 1989, Cox & Griggs 1982).

The same theoretical considerations provide a principled basis for predicting instances of transfer, or 'priming'. Although reasoning by analogy appears to be common in everyday life, it has been difficult to produce in the laboratory. The most common laboratory result is to find no transfer from a successfully solved problem to a target problem. But if one's model of a domain-specific inference mechanism is correct, then one should be able to reliably produce transfer to an ambiguous problem by first activating the appropriate mechanism with a clear instance of a problem drawn from that domain. Because this transfer is caused by the activation of a specific inference engine, this phenomenon can be called *inference priming*. Using the Wason selection task, we have been able to produce transfer to an ambiguous target problem in exactly this way (Fiddick et al 1995). Moreover, we have been able to prime social contract and precaution reasoning separately: i.e. we have been able to produce a double dissociation.

In these experiments, subjects were asked to solve two Wason selection tasks. The first — the prime — was either a clear social contract, a clear precaution rule or an ambiguous rule (as a control condition against which performance following the other two primes could be compared). In the second task — the target — the rule was always an ambiguous one that, when presented either alone or after another ambiguous rule, elicits low levels of performance. In some conditions, the target rule had the following two properties: (1) it could, in principle, be interpreted as a precaution rule; but (2) it would be difficult to interpret as a social contract. Let's call these 'ambiguous precaution rules'. An example would be 'if one empties the garbage cans, then one first eats red clay'. Emptying garbage cans and eating clay are both negative things — neither can be readily construed as a benefit. This should block a social contract interpretation. On the other hand, one could imagine situations in which emptying garbage could be hazardous (e.g. it could contain glass or disease-ridden materials), and sometimes people ingest substances to inoculate themselves from harm (e.g. penicillin). In other conditions, the target rule had opposite properties: (1) it could, in principle, be interpreted as a social contract; but (2) it would be difficult to interpret as a precaution rule. Let's call these 'ambiguous social contracts'. An example would be 'if one goes to the festival, then one lives in the village'. One can easily imagine situations in which going to the festival (or living in the village) could be rationed benefits, but neither of these activities sounds particularly hazardous.

We found the following. (1) A clear social contract strongly primed (i.e. elevated) performance on an ambiguous social contract target. Moreover, this was due to the activation of social contract categories, not logical ones: when the prime was a switched social contract, in which the correct cheater detection answer is not the logically correct answer (see Fig. 2), subjects' matched their answers on the target to the prime's benefit/requirement categories, not its logical categories. (2) A clear precaution rule strongly primed an ambiguous precaution target. Most importantly, (3) these effects were caused by the operation of two mechanisms, rather than one: a clear precaution rule produced little or no priming of an ambiguous social contract target; similarly, a clear social contract produced little or no priming of an ambiguous precaution target.

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

(4) Cross-cultural evidence

If an inference mechanism is part of the human cognitive architecture, then it should reliably develop in individuals (of a particular age/sex morph) across the ancestrally normal range of human environments. (Facultative adaptations would, of course, be an exception to this generalization; Williams 1966.)

According to Cheng & Holyoak (1985), the permission schema is not a component of the evolved architecture of the human mind. It is induced by content-independent

mechanisms of an unspecified kind, which are (presumably) evolved components of our computational architecture. The schema is induced when these mechanisms operate on information gleaned from the environment while the individual is attempting to achieve various goals. They do not specify what kind of information or goals are relevant; the implication is, however, that different cultural environments could lead to the induction of different schemas. The theory provides no a priori reason to expect the permission schema to be present in every human culture. Moreover, it suggests that the design of a schema should reflect the exigencies of life in the modern world, even if these bear no correspondence to the exigencies of life for ancestral hunter–gatherers.

In contrast, we have argued that reasoning about social contracts, precautions and threats is generated by three functionally distinct mechanisms; that each has a computational design that is specialized for solving the adaptive problems that typified their respective domains; and that each of these mechanisms is a component of the evolved architecture of the human mind — a reliably developing, species-typical set of cognitive procedures.

Statistical distribution of mechanisms. If our proposal is correct, then one should find manifestations of the same inference mechanisms across cultures. Moreover, their design features should reflect the statistical distribution of adaptive problems they evolved to solve. In contrast, structures built by content-independent mechanisms should reflect the statistical distribution of modern problems faced by the population under investigation. This is because the world experienced by an individual organism is the only source of content for a mechanism that starts out content free. For example, content-free learning mechanisms cannot account for the distribution of phobias: people rarely (if ever) develop phobias to electric sockets, cars and other dangers of the modern world. But they readily develop phobias of snakes and spiders — dangers faced by ancestral hunter–gatherers — even though these pose no significant danger in their personal lives (Marks 1987).

Analogously, the design and statistical distribution of reasoning mechanisms can provide evidence for or against alternative hypotheses about their genesis and structure. Even though the nature of the process that sculpts the permission schema is underspecified, one wonders why this process has not built reasoning mechanisms that are good at detecting violations of causal rules, for example. We live in a technological society. When an appliance stops working, we hypothesis test: the toaster heats up only if it is plugged in; is it plugged in? No, so it won't heat up ... and so on (Cosmides 1989).

Universality of mechanisms. Finding a culture in which the permission schema is absent would not count against that hypothesis. Finding a culture in which social contract algorithms were lacking would count against our hypothesis, however. Cognitive experiments supporting the hypothesis that there is a reasoning specialization for cheater detection have been conducted in different parts of the world (e.g. USA, UK, Germany, Hong Kong), but these sites were all in industrialized nations. Although each instance is informative, the evidence for species-typicality gains strength in proportion to the *diversity* of subject populations tested.

It is, of course, impossible to test for social contract algorithms in every human culture. An alternative is to test individuals from a culture that is different from our own along as many dimensions as possible. With Larry Sugiyama, we have been testing social contract reasoning among the Shiwiar, a population of hunter–horticulturalists in the rain forests of the Ecuadorian Amazon (Sugiyama et al 1995). The Shiwiar live in a culture that is about as different from industrialized society as currently exists, and which, in many ways, mirrors the kind of social environment in which humans evolved.

Shiwiar in our study area have no everyday direct contact with outsiders. They depend on hunting, fishing, gardening and foraging for their livelihood. Men continue to use traditional blowguns in hunting, although some use muzzle-loading shotguns as well when shot and powder is available. Relatively few Shiwiar speak Spanish. In day-to-day life Shiwiar is the dominant language, ties of kinship and affinity dominate social relationships mediated by gossip, witchcraft and the threat or use of violence, and distribution of goods is largely controlled through traditional systems of trade and kinship obligations. In short, although it is impossible to find a group of people who are not subject to some influence from the industrialized world, the Shiwiar in our study villages are at the far end of the spectrum in this regard. To the extent that they have been influenced by outsiders, it has been largely a material and not a psychological influence. Given that Shiwiar speak a non-western language, live in small isolated villages where they hunt, gather and practice swidden horticulture for their livelihood, and continue to interpret life with a Shiwiar world view, it would be difficult to argue that any convergence of experimental results between them and subjects in the industrialized world is due to western acculturation.

Sugiyama administered oral versions of the Wason selection task, testing social contract rules and descriptive rules. The Shiwiar tested showed the same pattern of responses as one finds in American college students.

(5) Developmental timetable

Ontogenetic evidence can also be used in dissecting cognitive architecture. In general, one expects cognitive adaptations to manifest a reasonably uniform timetable. Language acquisition unfolds in a uniform manner between 18 months and four years of age, for example (Pinker 1994); indeed, the acquisition of phonemes during the first year follows a uniform time-course whether they are being spoken by a hearing child or signed by a deaf child (Petitto & Marentette 1991). Uniform emergence is not, of course, a hard and fast rule: some adaptations mature in response to cues encountered by different individuals at diverse points in the life cycle and/or ones that some individuals never experience at all. Certain fish change sex in response to a social cue, for example. If a female blue-headed wrasse happens to be the largest in her group when the resident male dies, she turns into a male. Otherwise, she stays female (Warner et al 1975). But even in these cases, when one understands the design of the adaptation, one can predict its developmental timetable.

The same cannot be said for knowledge acquired through content-free inference procedures or as a by-product of adaptations designed to process information from other domains. Writing — a by-product of cognitive adaptations for language — is learned at many different ages and sometimes not at all. So are cooking, calculus and agricultural techniques. If they are induced via content-general procedures operating in goal-defined contexts, then there is no particular reason to expect the development of permission schemas (or social contract algorithms) to follow the same timetable across individuals or cultures.

So far, not much is known about the development of social contract algorithms and precaution rules. Preliminary evidence suggests, however, that they emerge fairly early. When given age-appropriate versions of the selection task, seven-year olds correctly detect violations of both social contracts and precaution rules (Girotto et al 1988); moreover, Cummins (1997) has found that three- and four-year olds correctly detect violations of precaution rules. (Children of this age have not yet been tested on rules that can be interpreted *only* as a social contract.) What is interesting about these data is the uniformity of emergence, not the early age at which this occurs — an adaptation can emerge at any point in the life cycle (e.g. beards, teeth, breasts).

Precocious performance is neither necessary nor sufficient for sustaining an adaptationist hypothesis. It is, however, relevant for evaluating claims of content-free learning (e.g. Markman 1989). The early age at which children solve these Wason tasks undermines the hypothesis that the domain-specific reasoning mechanisms responsible were constructed by content-independent procedures operating on individual experience. Pre-schoolers, who have a limited experience base, are not noted for the accuracy and consistency of their reasoning in many other domains, even ones with which they have considerable experience. For example, many children this age will say that a racoon can change into a skunk; that the word 'needle' is sharp; that there are more daisies than flowers; that the amount of liquid changes when poured from a short fat beaker into a tall thin one; and that they have a sister but their sister does not (Boden 1980, Carey 1984, Keil 1989, Piaget 1950). When a child has had experience in a number of domains, it is difficult to explain why a domain-general mechanism would cause the early and uniform acquisition of a reasoning schema for one domain yet fail to do so for others.

Conclusion

Dissecting computational architecture is, in essence, discovering its functional organization. This requires theories of function. Whether one is discussing human-made artefacts or biological systems, to characterize something as a *mechanism* is to commit oneself to the proposition that it has a given design because that design solves some problem. Cognitive scientists do not always acknowledge this, and so the assumptions about function that motivate their methods are sometimes left implicit. This does not always distort the study of information-processing adaptations: research on the eye has progressed, for example, because its function is so obvious and its design so closely parallels that of a machine designed by human engineers to solve a similar problem (the camera). But the eye is an exception. The function (if any) of most components of the human computational architecture is either unknown or so vaguely defined that hypotheses about their design cannot be motivated by reference to human-made machines. Absent a theory of function, there is no basis for deciding which machines are functionally analogous.

There are undoubtedly further methods that could be turned to the task of dissecting the architecture of our minds. But the human mind is a biological system, and the only process that creates functional organization in biological systems is natural selection. Methods rooted in the logic of adaptationism are the most efficient way to find that organization, because knowing the problem is halfway to knowing the solution.

Acknowledgements

For many clarifying discussions, we thank Clark Barrett, Don Brown, Larry Fiddick, Gerd Gigerenzer, Rob Kurzban, Steven Pinker, Melissa Rutherford, Roger Shepard, Valerie Stone, Larry Sugiyama and Don Symons. We also thank the James S. McDonnell Foundation and the National Science Foundation (NSF grant BNS9157–449 to J. T.) for their financial support.

References

Atran S 1990 The cognitive foundations of natural history. Cambridge University Press, New York

Axelrod R 1984 The evolution of cooperation. Basic Books, New York

Axelrod R, Hamilton WD 1981 The evolution of cooperation. Science 211:1390–1396

Baron-Cohen S 1995 Mindblindness: an essay on autism and theory of mind. MIT Press, Cambridge, MA

Boden M 1980 Jean Piaget. Viking, New York

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

Brown A 1990 Domain-specific principles affect learning and transfer in children. Cognit Sci 14:107–133

Bugental D, Goodnow J 1997 Socialization processes. In: Eisenberg N (ed) Handbook of child psychology: social, emotional and personality development. John Wiley & Sons Inc., New York

Carey S 1984 Cognitive development: the descriptive problem. In: Gazzaniga MS (ed) Handbook of cognitive neuroscience. Plenum, New York, p 37–66

Cheng P, Holyoak K 1985 Pragmatic reasoning schemas. Cognit Psychol 17:391–416

Cheng P, Holyoak K 1989 On the natural selection of reasoning theories. Cognition 33:285–313

Chomsky N 1957 Syntactic structures. Mouton, The Hague

Cosmides L 1985 Deduction or Darwinian algorithms? An explanation of the 'elusive' content effect on the Wason selection task. Doctoral dissertation, Harvard University, Cambridge, MA (University Microfilms #86-02206)

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

Cosmides L, Tooby J 1987 From evolution to behavior: evolutionary psychology as the missing link. In: Dupre J (ed) The latest on the best: essays on evolution and optimality. MIT Press, Cambridge, MA, p 277–306

Cosmides L, Tooby J 1989 Evolutionary psychology and the generation of culture. II. Case study: a computational theory of social exchange. Ethol Sociobiol 10:51–97

Cosmides L, Tooby J 1992 Cognitive adaptations for social exchange. In: Barkow J, Cosmides L, Tooby J (eds) The adapted mind. Oxford University Press, New York, p 163–228

Cosmides L, Tooby J 1994 Beyond intuition and instinct blindness: toward an evolutionarily rigorous cognitive science. Cognition 50:41–77

Cox J, Griggs R 1982 The effects of experience on performance in Wason's selection task. Mem Cognit 10:496–502

Cummins D 1997 Evidence of deontic reasoning in 3- and 4-year old children. Mem Cognit, in press

Dawkins R 1982 The extended phenotype. Freeman, San Francisco, CA

Dawkins R 1986 The blind watchmaker. Norton, New York

Ekman P 1992 Facial expressions of emotion: new findings, new questions. Psychol Sci 1:34N–38N

Etcoff N 1984 Selective attention to facial identity and facial emotion. Neuropsychologia 22:281–295

Etcoff N, Freeman R, Cave K 1991 Can we lose memories of faces? Content specificity and awareness in a prosopagnosic. J Cognit Neurosci 3:25–41

Fernald A 1992 Human maternal vocalizations to infants as biologically relevant signals: an evolutionary perspective. In: Barkow J, Cosmides L, Tooby J (eds) The adapted mind. Oxford University Press, New York, p 391–428

Fiddick L, Cosmides L, Tooby J 1995 Priming Darwinian algorithms: converging lines of evidence for domain-specific inference modules. Seventh annual meeting of the Human Behavior and Evolution Society. University of California, Santa Barbara, CA

Fiske A 1991 Structures of social life: the four elementary forms of human relations. Free Press, New York

Fodor J 1983 The modularity of mind: an essay on faculty psychology. MIT Press, Cambridge, MA

Frith U 1989 Autism: explaining the enigma. Blackwell, Oxford

Gigerenzer G, Hug K 1992 Domain-specific reasoning: social contracts, cheating and perspective change. Cognition 43:127–171

Girotto V, Light P, Colbourn C 1988 Pragmatic schemas and conditional reasoning in children. Q J Exp Psychol Sect A Hum Exp Psychol 40:469–482

Gopnik M 1990 Dysphasia in an extended family. Nature 344:715

Hatano G, Inagaki K 1994 Young children's naive theory of biology. Cognition 50:171–188

Jackendoff R 1992 Languages of the mind. MIT Press, Cambridge, MA

Johnson-Laird P, Byrne R 1991 Deduction. Lawrence Erlbaum Associates Inc., Hillsdale, NJ

Kahneman D, Slovic P, Tversky A 1982 Judgment under uncertainty: heuristics and biases. Cambridge University Press, New York

Keil F 1989 Concepts kinds and cognitive development. MIT Press, Cambridge, MA

Keil F 1994 The birth and nurturance of concepts by domain: the origins of concepts of living things. In: Hirschfeld L, Gelman S (eds) Mapping the mind: domain specificity in cognition and culture. Cambridge University Press, New York, p 234–254

Leslie A 1987 Pretense and representation: the origins of 'theory of mind'. Psychol Rev 94:412–426

Leslie A 1988 Some implications of pretense for the development of theories of mind. In: Astington JW, Harris PL, Olson DR (eds) Developing theories of mind. Cambridge University Press, New York, p 19–46

Leslie A, Thaiss L 1992 Domain specificity in conceptual development: neuropsychological evidence from autism. Cognition 43:225–251

Lewontin R 1979 Sociobiology as an adaptationist program. Behav Sci 24:5–14

Maljković V 1987 Reasoning in evolutionarily important domains and schizophrenia: dissociation between content-dependent and content independent reasoning. Undergraduate honors thesis, Harvard University, Cambridge, MA, USA

Manktelow K, Over D 1990 Deontic thought and the selection task. In: Gilhooly KJ, Keane MTG, Logie RH, Erdos G (eds) Lines of thinking, vol 1. Wiley, Chichester, p 153–164

Mann J 1992 Nurturance or negligence: maternal psychology and behavior preference among preterm twins. In: Barkow J, Cosmides L, Tooby J (eds) The adapted mind. Oxford University Press, New York, p 367–390

Markman E 1989 Categorization and naming in children. MIT Press, Cambridge, MA

Marks I 1987 Fears phobias and rituals. Oxford University Press, New York

Petitto L, Marentette P 1991 Babbling in the manual mode: evidence for the ontogeny of language. Science 251:1493–1496

Piaget J 1950 The psychology of intelligence. Harcourt, New York

Pinker S 1994 The language instinct. Harper Collins, New York

Platt R, Griggs R 1993 Darwinian algorithms and the Wason selection task: a factorial analysis of social contract selection task problems. Cognition 48:163–192

Rips L 1994 The psychology of proof. MIT Press, Cambridge, MA

Spelke E 1990 Principles of object perception. Cognit Sci 14:29–56

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

Sugiyama L, Tooby J, Cosmides L 1995 Cross-cultural evidence of cognitive adaptations for social exchange among the Shiwiar of Ecuadorian Amazonia. Seventh annual meeting of the Human Behavior and Evolution Society. University of California, Santa Barbara, CA

Tooby J, Cosmides L 1990 On the universality of human nature and the uniqueness of the individual: the role of genetics and adaptation. J Pers 58:17–67

Tooby J, Cosmides L 1992 The psychological foundations of culture. In: Barkow J, Cosmides L, Tooby J (eds) The adapted mind. Oxford University Press, New York, p 19–136

Tooby J, Cosmides L 1997 Ecological rationality in a multimodular mind. In: Cummins D, Allen C (eds) The evolution of mind. Oxford University Press, New York, in press

Trivers R 1971 The evolution of reciprocal altruism. Q Rev Biol 46:35–57

Warner R, Robertson D, Leigh E 1975 Sex change and sexual selection. Science 190:633–638

Wason P 1983 Realism and rationality in the selection task. In: Evans J St BT (ed) Thinking and reasoning: psychological approaches. Routledge, London, p 44–75

Wason P, Johnson-Laird P 1972 The psychology of reasoning: structure and content. Harvard University Press, Cambridge, MA

Williams G 1966 Adaptation and natural selection. Princeton University Press, Princeton, NJ

Wynn K 1992 Addition and subtraction by human infants. Nature 358:749–750

DISCUSSION

Buss: You have extracted the social contract from the natural relationships in which they evolved. For example, mateships and friendships are two different types of relationship in which social contracts come into play, and what constitutes cheating

is different in those two relationships: having sex with someone outside a relationship would be constituted as cheating or a violation of a contract but this wouldn't be the case in the context of a friendship, whereas failure to reciprocate immediately within a friendship may be a violation but this may not be the case in a mateship. How can you deal with the issue of even more domain specificity than you're been arguing for?

Cosmides: First one needs to develop a theory of the adaptive problems involved in mating and friendship. At that point one can ask, like an engineer, 'What properties would we expect a mechanism well designed for detecting cheating in mateships to have? Are these the *same* properties we would expect of a mechanism well designed for detecting cheating in friendships? Could one, more general, social contract mechanism solve both adaptive problems? Or should we expect to find one mechanism that deals with social contracts outside the domain of mateships, and another specialized for mating relationships?' If we decided that there might be a mechanism specialized for detecting cheating in mateships, then we could conduct reasoning experiments transforming content and context, and see, empirically, whether there is a syntax specialized for that adaptive domain. Developing the theory is the crucial step, and it can reveal dimensions of an adaptive problem that common sense notions of 'reciprocity' or 'cheating' would not. For example, John Tooby and I have been working on a model of selection pressures based on what we call the 'Banker's Paradox'. The model suggests that friendship is not based on reciprocation, but rather on a form of deep engagement that has more in common with economic models of insurance. This leads to different psychological predictions. For example, if mechanisms that govern friendship were shaped by selection pressures for reciprocity, then people should pick friends who have *different* tastes than themselves, because this provides more opportunities for gains in trade. In contrast, the Banker's Paradox model predicts you will pick friends with tastes similar to your own; moreover, it predicts that reciprocating immediately will be construed as a sign that you are not a person's friend.

Buss: But in some social contracts, such as mateships, the nature of the contract is to tag to specific types of commodities, such as sexual commodities, and the way you develop the theory is content independent. A complete theory of social contracts would have to specify the nature of the content of the exchange, which may differ from species to species and from relationship to relationship.

Cosmides: I doubt that the social exchange mechanisms that I have been describing apply to the mating domain at all. (Indeed, excessive concern with reciprocation is probably the death knell of a mateship.) In my view, there is probably a separate set of mechanisms that govern reasoning, decision making, partner evaluation, preferences and memory for information involving mating than for other kinds of relationships — i.e. there are probably mechanisms that are specialized for mateships.

Nisbett: I have a comment on the opposite topic. David Buss is talking about highly specific domains, whereas I believe there are also general inferential rules, some of which I would argue we share with animals. For example, we seem to share temporal considerations for assessing causality: the ability to learn an association between two arbitrary stimuli doesn't seem to last beyond a few seconds. Holyoak et al (1989) have argued for an 'unusualness heuristic'. For decades it was assumed that many trials were necessary before animals could learn associations between arbitrary events. But Leon Kamin showed that one-trial learning was possible. Holyoak et al (1989) argued that a highly general heuristic can account for such results. That is, organisms will form an association when one unexpected event is followed by another unexpected event. This unusualness heuristic is absolutely content free.

Cosmides: I'm not sure that it is necessarily content free. There is domain specificity in what gets defined as unexpected versus not unexpected. Therefore, in the Garcia & Koelling (1966) experiments, if rats are nauseated after eating food they infer that it is the food that is nauseating them, but if they are nauseated after seeing a red light they don't infer that the red light is causing this.

Nisbett: This is the opposite of what we've been talking about: this is clearly domain specific.

Cosmides: I'm saying that a red light followed by nausea is an unexpected event, but they do not become conditioned by it.

Nisbett: It's true that a red light followed by nausea would not be expected, but it would never be noticed either because you can't produce instant nausea. Even if you could, you would probably not have one-trial learning, as you do with novel tastes. Organisms are 'counter prepared' to learn associations such as those between light and illness, whereas they are so well prepared to learn associations between novel foods and nausea that you can have a gap between the food and the nausea of 24 h or more and still get learning. In contrast, the maximum gap for arbitrary associations—that is, those for which the organism is neither prepared nor counter-prepared—is a matter of seconds. What Holyoak et al (1989) are referring to is these non-prepared associations for which it is possible to build up expectations within the context of the experiment. For example, consider the situation in which a rat learns over many trials that a buzzer signals imminent shock, and then has a trial in which a bright light immediately precedes the buzzer but the buzzer is not followed by shock. On the very next occasion on which the light precedes the buzzer, the rat acts as if it 'holds the hypothesis' that the buzzer does not signal shock as usual. Thus, the unusual event of the bright light preceding the buzzer is somehow linked in the rat's mind to the unusual event of the buzzer not predicting shock. Moreover, within just a few trials, the rat is at asymptote for learning. This is not like standard Skinnerian learning, in which many pairings are necessary to reach asymptote. Hence, Holyoak et al (1989) maintain that the rat has a heuristic it is applying rather than relying on mere brute associative learning.

Other general inferential rules are probably limited to humans. One example of this is the sunk-cost rule, which describes the tendency for people to consume something just because they have paid for it. You can teach people that they're following this rule by making them realize that consuming something with negative value is not a good idea, and you can change their behaviour across a wide range of domains. It's difficult to imagine that this is an inferential rule that we share with animals.

Tooby: We're not arguing that there are no general rules. We are just suggesting that psychologists should consider the hypothesis that a given performance is generated by domain-specific mechanisms on an equal basis with the hypothesis that it is generated by domain-general mechanisms, rather than either ruling it out a priori or accepting a lower standard of evidence for the domain-general hypothesis.

Cosmides: Adaptationism can help one understand when a cognitive system will be relatively content independent as well. For example, there are certain kinds of problems where you have to track the frequency of events ontogenetically, so that the phylogenetic specification of those frequencies won't really help: one can't specify phylogenetically that there's going to be more elk in one canyon versus another canyon. So different, more content-general mechanisms exist to track that kind of information. So far, the only limitation on the frequency computation system that we have found is the ability to individuate an object. For example, if you give somebody something that they don't token individuate, such as the sides of cards, they do badly on probabilistic reasoning tasks.

Hauser: I would like to raise a comment about our inability to detect deception. Have you tried running these kinds of experiments with logicians, who presumably have no problem with conditionals? If they showed a faster response to a social contract than a standard modus tolens problem wouldn't that be even stronger evidence for your claim about specialization?

Cosmides: That's an interesting idea. I haven't yet tried that, although logicians had difficulty with the original abstract selection test. There are some differences in performance between populations, which I suspect are due to differences in the ability to solve pencil-and-paper problems. If you give German undergraduates the same battery of problems, the difference in performance between social contracts and descriptive problems is identical to when they are given to American undergraduates, but their scores are shifted up slightly. These studies were performed by Gerd Gigerenzer and it's interesting that only 6% of subjects gave the logically correct answer on every single problem, even when the logically correct answer would fail to detect cheaters (Gigerenzer & Hug 1992). When Gerd asked them about this in the debriefing they all said that they solved the problems by applying the rules of inference of the propositional calculus. Furthermore, they said it was difficult to do this for some of the problems: it turned out that those were the switched social contracts, where the correct answer for detecting cheaters is not the logically correct answer.

Hauser: Has anyone timed how long it takes them to give the correct answer?

Cosmides: No.

Tooby: In many of these experiments we are tracking the proper adaptive responses. Your prediction that logicians would do this faster may not necessarily hold true. Logicians would, because of their training, have the tendency to give the answer that is correct from the point of view of formal logic, which in many cases would be an inappropriate response, rather than the answer that is correct from an adaptive point of view.

Gigerenzer: Logicians usually don't believe that everything can be reduced to one form of logic, such as first-order logic. Some try to develop domain-specific forms of logic. These logicians wouldn't even think of treating conditionals in a content-independent way.

I would like to raise a more general point. If psychological adaptations are, to some important degree, modular then we need to ask, how should we think of a module? The examples we have heard suggest two ways. The first is to assume that modules are psychological adaptations designed to solve important adaptive problems, such as a module for mate choice and one for social contracts. For instance, Leda Cosmides and John Tooby's social contract module is of that type — a module that integrates a specific mix of tools, including mechanisms for face or voice recognition, cheater detection, and characteristic emotions and behaviours. A second way to think about a module is not in terms of a characteristic mix of tools but in terms of a single tool. Proposals such as that of a number module, or a language module, seem to fall under this second view.

Cosmides: There are two different ways of thinking about this. The first addresses the question of whether the visual system is just one adaptation or a collection of adaptations, the outputs of which are functionally integrated to produce scene analysis. Scene analysis, i.e. knowing what things are present in the world and where they are, is just as much an adaptive problem as social exchange, but solving that problem supports many other activities. The same is true of language. For example, if you have language then you can make contracts for the future. One example that illustrates the interrelationship between adaptations is the theory that autism is a selective impairment of a theory of mind mechanism (Baron-Cohen et al 1985). There are different components of social exchange, and some clearly need a theory of mind. For example, if I am going to offer you something that you may want I have to be able to model what your desires are. But to detect cheating, it's not clear whether you can't just look for certain behavioural events. Similarly, it is not clear how much 'general intelligence' one needs for reasoning about social exchange. Maljković (1987) showed that people with schizophrenia have impaired general reasoning, but their ability to detect cheaters is intact. Finding double dissociations can help clarify these questions. For example, people with autism may have a selectively impaired theory of mind but they can have normal IQs, whereas people with Williams' syndrome seem to have an intact theory of mind but can be profoundly retarded and do badly in spatial tests. One can ask, will a person with autism have trouble with social exchange but be good at detecting violations of precaution rules, which are not social rules? Would we find the opposite in people with Williams' syndrome, i.e. that they are good at detecting cheaters but not at detecting rotations of precaution rules? Ultimately, it will be interesting to find out how these are related and whether the output of one mechanism provides input to another mechanism.

References

Baron-Cohen S, Leslie A, Frith U 1985 Does the autistic child have a 'theory of mind'? Cognition 21:37–46

Garcia J, Koelling R 1966 Relations of cue to consequence in avoidance learning. Psychon Sci 4:123–124

Gigerenzer G, Hug K 1992 Domain-specific reasoning: social contracts, cheating and perspective change. Cognition 43:127–171

Holyoak KJ, Koh K, Nisbett RE 1989 A theory of conditioning: inductive learning within rule-based default hierarchies. Psychol Rev 96:315–340

Maljković V 1987 Reasoning in evolutionarily important domains and schizophrenia: dissociation between content-dependent and content-independent reasoning. Undergraduate honours thesis, Harvard University, Cambridge, MA, USA