

A Logical Design for the Mind?

Lance J. Rips
The Psychology of Proof: Deductive Reasoning in Human Thinking
Cambridge, MA: MIT Press, 1994.
449 pp. ISBN 0-262-18153-3. \$45.00

Review by
Leda Cosmides and John Tooby

Lance J. Rips, professor of psychology at Northwestern University (Evanston, Illinois), is author of the chapter "Deduction and Cognition" in D. N. Osherson (Ed.) *An Invitation to Cognitive Science*. ■ Leda Cosmides and John Tooby are associate professors in the Psychology and Anthropology Departments of the University of California, Santa Barbara, where they are codirectors of the Center for Evolutionary Psychology. They are coeditors, with J. Barkow, of *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Cosmides received the 1988 American Association for the Advancement of Science Prize for Behavioral Science Research for her work on human reasoning. She received a 1993 American Psychological Association Early Career Award, and Tooby a 1991 National Science Foundation Presidential Young Investigator Award, for their contributions to evolutionary psychology.

This century has witnessed a florescence in the development of formal systems and methods of all kinds—in mathematics, probability theory, logic, linguistics, artificial intelligence, game theory, economics, and so on. It is a curious feature of the cognitive revolution that, out of all of these formal approaches, modern logic has barely been mined by cognitive psychologists as a source for theories of cognitive processes. This is particularly odd given the impressive power of modern logical systems as automatable knowledge-productive and truth-preserving systems.

One reason for this failure to exploit much of the rich array of extant logics is that most cognitive psychologists came to the conclusion that human reasoning does not operate according to a mental logic. That is, thinking in general and reasoning in particular is not widely believed to consist of the application of implementations of logical rules to representations that are encoded on the basis of their abstract logical forms. Although reasoning research began with some expectation that the rules of thought might well be the rules of logic, a series of experiments showed that subjects' performance on many logical tasks was quite poor—tasks that would be trivially easy if humans had simple deductive procedures online (e.g., Wason & Johnson-Laird, 1972). This led to the widespread rejection of "the" mental logic hypothesis. This conclusion was further solidified by a flood of results from the decision-making community—results that have been taken to be demonstrations of pervasive human irrationalities and sus-

ceptibilities to fallacies in both inductive and deductive reasoning (Kahneman, Slovic, & Tversky, 1982). Because it seemed that even many simple logical steps were beyond the reach of untutored subjects, the majority of cognitive psychologists decided logic was not relevant to fundamental psychological processes and thus have not bothered to acquaint themselves with advances in modern logic and the rich multitude of alternative logics and methods that are now available. For many, the search for models of how human reasoning actually occurs has turned instead to alternative approaches such as heuristics, schemas, or mental models—candidate processes that (allegedly) do not require the existence of rule-based deductive abilities to operate.

Against this background, Lance Rips's new book, *The Psychology of Proof: Deductive Reasoning in Human Thinking*, stands out as the most substantial and sophisticated defense of the mental logic position in the history of the field. In this book, Rips advances what he calls the deduction-system hypothesis: that all normal members of the human species are equipped with an innate and extensive set of deduction principles that govern how we reason, and that this natural deduction system forms the core of many—perhaps most—cognitive activities. Although he intends his theory to account for human deductive reasoning, he also argues that these deduction procedures may be, in effect, the "general-purpose programming" (p. viii) language of the human cognitive architecture within which a variety of nondeductive cognitive processes could also be imple-

mented. Indeed, one can begin to appreciate the computational power inherent in deduction systems by recognizing that widely favored alternatives such as Anderson's ACT* (1983) and Newell's Soar (1990) are special cases of the deduction-system architecture Rips develops, versions that are computationally crippled by being limited to just two of the many rules available to Rips's architecture (modus ponens and universal instantiation). In contrast to the vague handwaving that has become an unfortunate feature of many cognitive theories, Rips is admirably explicit about the exact form of the procedures he is positing. His theory, PSYCOP (short for psychology of proof), exists as a computer implementation, allowing him to make complex yet constrained empirical predictions through simulations of how subjects would reason their way through intricate arguments. Moreover, the self-imposed discipline of turning the formal system into a set of computer procedures prompted important developments in Rips's model, by sharpening awareness of a variety of critical computational issues (e.g., combinatorial explosion, process control) that, in other hands, are too often swept under the rug. One of the most appealing features of Rips's thinking and model is the originality with which he selected and integrated plausible ideas from a broad variety of research literatures, including recently developed techniques in logic, methods derived from automatic theorem-proving research in artificial intelligence, and, of course, the body of research that has accumulated in the study of the psychology of reasoning.

Rips is successful in his primary goal of restoring the credibility of the mental logic hypothesis as a viable candidate account of human reasoning, and perhaps also as a candidate cognitive architecture on a rough par with production-based and connectionist systems. Disinterested cognitive scientists would, we suspect, judge that this book persuasively shows that most of the arguments that have been used to dismiss the mental logic hypothesis are either unfounded or weigh as strongly against rival approaches as they do against the mental logic position. Rips's formal model of how mental proofs might operate compares favorably, both theoretically and empirically, with its primary "general-purpose" theoretical competitor, the mental models approach (Johnson-Laird & Byrne, 1991). Indeed, one of the most incisive sections of the book is Rips's dissection of the problems

with, and unfounded claims made for, mental models theory. Anyone who has tried to construct experiments to test between mental models theory and alternative reasoning hypotheses will sympathize with Rips's assessment that Johnson-Laird and Byrne's theory "is inexplicit in ways that make it difficult to understand the basis of their predictions" (p. 359). Most of what is necessary to make mental models theory predict outcomes in a tightly constrained manner is passed off to black boxes such as procedural semantics and world knowledge. It is to Rips's credit that, despite this obstacle, he still finds some serviceable ways to empirically contrast his deduction system with mental models analyses.

So what significant features of human psychology would a mental logic plausibly account for? Some mental activity does seem to consist of sentence-like representations that, through the application of various procedures, produces additional propositions, often in an ordered chain. These additional propositions are then treated by the cognitive architecture as true or provisionally true. Many of these apparently direct (i.e., one-step) sentence transformations or derivations seem to correspond to at least some of the logical operations postulated in canonical logical systems. Many basic concepts used in everyday thought and language also correspond, in some measure, to logical primitives such as truth, negation, contradiction, conjunction, various conditionals, quantifiers, and so on. The fact that humans everywhere routinely reason with suppositions strongly argues that, if our minds embody a single deduction system, it is a natural deduction system rather than an axiomatic one. (This capacity to entertain a proposition provisionally in order to follow out its consequences is one of the marked parallels between human reasoning and natural deduction systems.) Moreover, it is telling that one factor that inclines humans to accept or reject arguments seems to be whether their steps conform to or violate certain logical operations. Subjects can make judgments of inferential soundness that seem analogous to judgments of grammaticality, and they can judge or generate inferences for a potentially infinite set of instances (e.g., consider the logical rule *modus ponens*: If A then B. A. Therefore B.—where an infinite set of instances can be substituted for A and B). This last fact alone implies, Chomsky fashion, that humans must be able to encode and operate on propositions at an extremely abstract

level, and it is tempting to believe that the properties of the operations available at this abstract cognitive level are the source of the intuitions that have been formalized by logicians into logical systems.

Although these evident facts seem to be better explained by the mental logic position than by alternatives, there are a series of difficulties with the mental logic position that originally led it into disfavor and that still make it hard for many to accept. First of all, although untutored subjects do seem to be able to engage in deductive reasoning with a far better than random proficiency (a fact well explained by the mental logic hypothesis), their performance is also often very bad, and seems quite remote from the proficiency one might expect from a straightforward mechanical implementation of most standard logical systems. Rips properly insists that one must have a principled explanation for the existence of good performance by subjects on deduction tasks and makes a plausible case that these successes are hard to explain in the absence of a deduction system incorporating at least some of the features of the one he proposes. However, his account of reasoning failures involves some assumptions that seem very peculiar. In his interpretation of errors, Rips makes the usual competence-performance distinctions, and also makes some standard assumptions about working memory limitations and similar factors that can be assumed to degrade performance. However, to reconcile the high rate of subject errors with the presence of a large battery of innate logical operations, Rips is forced to assume that these procedures are randomly unavailable to the architecture—and at extraordinarily high rates. Rips uses parameter fitting on various data sets to derive estimates of the average availability of various rules (assuming his model to be true) and although some rules, such as *AND Introduction* and *Disjunctive Modus Ponens* appear to be always available to the reasoning system, others, such as *OR Introduction* are randomly unavailable over 80 percent of the time. This is a strange design for what is, ultimately, a system for generating adaptive behavior. After all, the key virtue of a rule-based as opposed to, say, a connectionist architecture is its computational sensitivity: Its output can vary radically depending on exactly which rules get applied to a given set of representations. But what kind of engineering sense does it make to design a behavioral control system whose deci-

sions on identical tasks vary wildly because many of its component rules are randomly and routinely unavailable? Most real-world tasks require multiple reasoning steps, and with Rips's estimates of the probabilities of rule unavailabilities, the multiplication of conditional probabilities over steps would lead to very steep performance degradation as reasoning sequences become longer than two or three steps. Any deduction system designed in this way would deliver strikingly erratic judgments as rules come randomly on and off-line. Researchers may differ in how plausible they consider such a design to be, but it seems illuminating to contrast subjects' failure rates on simple reasoning problems with human grammatical competence. Grammatical assignments in speech processing seem to require at least as many inferences as the reasoning experiments Rips analyzes, and yet humans successfully perform these tasks with extremely high reliability. As implausible as this aspect of Rips's design seems to us, however, it is worth mentioning that he is able to account broadly for some interesting properties of these data sets, such as problem difficulty, by categorizing reasoning problems on the basis of which rules of deduction they require for their solution. A mental logic position seems to explain some features of reasoning successes yet seems unsatisfying in its account of reasoning errors.

A second notable difficulty with the mental logic position is that any system that operates on representations solely on the basis of their logical form ought to be relatively insensitive to their content. Yet, human reasoning is notoriously sensitive to content, and content effects often dominate experimental outcomes. (The power of these content effects in reasoning has motivated the construction of a new generation of reasoning theories that are domain-specific, such as Cheng & Holyoak's pragmatic reasoning schemas [1985] or our and Gigerenzer's work on social contracts [e.g., Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992].) A mental logic approach cannot easily or naturally be made to explain content effects because of the intrinsic content independence of its procedures—although as Rips carefully points out, a mental models approach is not equipped to do any better. To account for content effects in reasoning, content-independent reasoning theories must rely on (a) problem-specific assumptions about what additional propositions subjects introduce into their rea-

soning that are not given to them declaratively by the experimenter, (b) assumptions about the nature of the interpretive apparatus that maps the explicit problem content given to the subject onto the internal representations that are then acted on by the reasoning procedures present in the subject's mind, and (c) the assumption that a particular set of rules will reliably be activated by representations with this content. To the best of our knowledge, no content-independent reasoning theory has any principled theory or account of these crucial elements. This prevents them from making precise predictions about how subjects will reason about individual problems on the basis of their content. For this reason, explanations of content effects by advocates of general purpose reasoning theories are inevitably post hoc—something Rips is frank about.

Rips is also quite aware of the need for a principled theory of how subjects interpret reasoning problems, but cannot supply one because his goal is to provide a global account of reasoning that covers all potential contents. Such a global reasoning system would need to be paired with an interpretive apparatus that could interpret all contents, which is nothing less than a cognitive model of adult encyclopedic knowledge and discourse analysis. A way around this roadblock is to develop theories of interpretation and inferential competence for circumscribed domains that make explicit predictions about exactly what additional entailments subjects will make in response to particular problem contents—the approach we took with the development of social contract theory, and the analysis of threats (see, e.g., Cosmides & Tooby, 1989, 1992; Tooby & Cosmides, 1989). Rips is resistant to these approaches because they seem to be unparsimonious—multiplying mental entities where perhaps one comprehensive general purpose system would serve.

So, the field of reasoning research is left with an apparent contradiction: As Rips's book argues, certain facts about human reasoning seem almost to demand the existence of a mental logic—yet accepting Rips's hypothesis that the mind operates according to a mental logic leads to other predictions that seem to be contradicted by other apparent facts. There may be several paths out of this predicament, but we are aware of only one: to accept, as Rips suggests, that our reasoning faculties include some procedures that map onto logical rules such as modus ponens, AND Elimination, universal in-

stantiation, and so on; but also to accept what Rips rejects—that our reasoning faculties simultaneously include a battery of specialized inference procedures that respond not to logical form but to content types, procedures that do not correspond to logical rules but that coexist in the same system with them.

This might be termed the ecological rationality position (see Cosmides & Tooby, 1992; Gigerenzer, 1991) because it posits inferential procedures that were designed (by evolution or ontogenetic calibration) to work well within the stable ecological structure of the domain they were designed to operate on, even though they might lead to false or contradictory inferences if they were activated outside of that domain. Such a position seems to offer an escape from the apparent contradictions within the reasoning literature, and a series of other problems as well. It would explain how humans can often make successful deduction-like arguments, and yet systematically depart from logical rules in predictable content-sensitive ways. It would explain how humans can be so competent at solving natural problems (specialized inference procedures guide them through a stably structured world in ways that are more effective than generalized techniques could), and yet seem to be so error prone on experiments: When the standard of error is deviation from general purpose deductive or inductive logics, the operation of specialized rational methods will seem erroneous (Gigerenzer, 1991). It would explain how developing children could bootstrap their way into a rich encyclopedic knowledge of the world—something content-independent reasoning systems are incapable of doing unassisted. Indeed, it would help to integrate the mental logic position with the recent advances in cognitive development that indicate that infants come equipped with just such domain-specialized inferential principles (Baron-Cohen, 1995; Hirschfeld & Gelman, 1994). Whichever of the current theoretical positions (mental models, mental logic, ecological rationality, etc.) turns out to be close to the truth (if any does), it will owe a debt to Rips—either by specifying procedures that the mind is likely to include, or by challenging rival theories to provide better explanations for the deduction-like performances humans seem capable of generating.

References

Anderson, J. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

- Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.
- Cheng, P., & Holyoak, K. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391–416.
- Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Experiments with the Wason selection task. *Cognition*, 31, 187–276.
- Cosmides, L., & Tooby, J. (1989). Evolutionary psychology and the generation of culture, Part II. Case study: A computational theory of social exchange. *Ethology and Sociobiology*, 10, 51–97.
- Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology*, 2, 83–115.
- Gigerenzer, G., & Hug, K. (1992). Domain specific reasoning: Social contracts, cheating and perspective change. *Cognition*, 43, 127–171.
- Hirschfeld, L., & Gelman, S. (1994). *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University Press.
- Johnson-Laird, P., & Byrne, R. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Tooby, J., & Cosmides, L. (August, 1989). The logic of threat: Evidence for another cognitive adaptation? Paper presented at the *Human Behavior and Evolution Society*, Evanston, IL.
- Wason, P., & Johnson-Laird, P. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.