Rules and Illusions: A Critical Study of Rips's *The Psychology of Proof*

PHILIP N. JOHNSON-LAIRD
Department of Psychology, Princeton University, Princeton, NJ 08544, U.S.A.
philclarity.princeton.edu

Lance J. Rips, *The Psychology of Proof: Deductive Reasoning in Human Thinking*, Cambridge, MA: MIT Press, 1994, xiii + 449 pp., $45.00 (cloth), ISBN 0–262–18153–3.

> If Rips is right, there are formal rules of inference in the mind;
> *or else* if Rips is wrong, there are formal rules of inference in the mind.

Human reasoning is a mystery. Is it at the core of the mind, or an accidental and peripheral property? Does it depend on a unitary system, or on a set of disparate modules that somehow get along together to enable us to make valid inferences? And how is deductive ability acquired? Is it constructed from mental operations, as Piagetians propose; is it induced from examples, as connectionists claim; or is it innate, as philosophers and "evolutionary psychologists" sometimes argue? Is deduction a matter of mobilizing formal rules of inference like those of a logical calculus, or of rules with a specific content like those of a computer "expert system", or of remembered cases of valid reasoning like those exploited in other AI programs? Or could it depend on a grasp of meaning and of the fundamental semantic principle that a conclusion is valid if there are no cases in which the premises are true but it is false? Psychologists have been struggling with deduction for a century; cognitive scientists have recently honed in on it, and they have proposed explicit "information-processing" models of the process. Each of the positions in the list above has its defenders, and the controversy is hot.

The Psychology of Proof presents a comprehensive theory that the mind is equipped with formal rules of inference. Lance Rips published an initial theory in 1983, and the present account is his *summa theologicum*. It defends deduction as a central cognitive ability; it defends formal rules as the basic symbol-manipulating operators of cognitive architecture; and it defends formal rules as the lower-level principles that guide deductive thinking. It describes a set of rules that for the first time accommodate reasoning with sentential connectives (such as *if*, *and*, and *or*) and quantifiers (such as *all* and *some*) within a psychological theory. Rips calls the system PSYCOP – nothing to do with the "thought police", but an acronym from *psyc*hology *o*f *p*roof – and he has both implemented it in Prolog and tested it experimentally with some success. The book is a major achievement, and it

Minds and Machines **7:** 387–407, 1997.

should be read by anyone interested in how people reason, though it is technically demanding. Part I reviews the psychology of reasoning, formal logic, and automated theorem-proving. Part II describes PSYCOP and assesses the evidence in its favor. Part III considers other sorts of reasoning and other sorts of theories of reasoning – alternative formal-rule theories, theories based on productions or pragmatic schemas, and theories based on mental models. It makes some cogent points against them and argues that PSYCOP has advantages over all its rivals.

At this point, I should declare an interest. Although, at one time, I too argued that the mind might be equipped with formal rules of inference (Johnson-Laird 1975), I also suggested in the same paper that reasoning might be based on mental models of the states of affairs described by premises – a view that now seems to me to give a better account of human reasoning than theories based on formal rules of inference (see Johnson-Laird 1983, Johnson-Laird and Byrne 1991). Hence, I should say at the outset: I admire Rips's book, but I do not accept its basic argument. My plan in what follows is, first, to outline Rips's Deduction-System hypothesis; second, to describe PSYCOP in sufficient detail for it to be understood by newcomers, exposing some flaws along the way – flaws that for the most part can be fixed; third, to consider the evidence in favor of the theory; and, finally, to address the viability of the enterprise as a whole, touching upon evidence that strikes at its foundations.

1. Deduction-System Hypothesis

The paradigm of a formal rule of inference is *modus ponens*, which sanctions inferences of the form:

If P then Q
P
∴ Q.

Rips begins with the idea that formal rules of inference, such as *modus ponens*, are central to human cognition, underlying not just deduction but thinking in general. He calls this idea "the Deduction-System hypothesis". It implies that formal rules are part of cognitive architecture and that they constitute a system akin to a general-purpose programming system. Developing and testing the Deduction-System hypothesis, Rips tells us, is the main goal of his book.

One critic of the use of logic as a psychological theory is my Princeton colleague, the philosopher Gilbert Harman. As he points out, logic is an account of the implications between sets of sentences in a formal language, whereas reasoning is a mental process that affects beliefs (Harman 1986). Suppose, for instance, that you believe the following two propositions:

> If the epigraph of this review means what it seems to say, then Phil believes that formal rules of inference underlie reasoning.

and:

The epigraph of this review does mean what it seems to say.

You can accordingly deduce:

Phil believes that formal rules of inference underlie reasoning.

Alas, as you read on, you will see that it would be folly to believe this conclusion. Something has to give, and what gives, presumably, is your belief in one or other of the premises. A theory of reasoning, Harman maintains, should account for this change in belief, and logic alone is impotent to explain how you change your mind. What is needed is a theory of how inferences lead to the best explanation of phenomena, and formal rules of deductive inference may not have any privileged status in such a theory.

Another way of making the same point is that human reasoners make inductions that go beyond the information that is given to them. I park my Rolls within the city walls of Siena, and the police tow it. I infer: If a tourist parks within the city walls of Siena, then the police will tow the car. Such inferences are commonplace, though they are not deductively valid. Some are stronger than others, but their strength cannot be accounted for in terms of deductive rules (see Osherson, Smith, and Shafir 1986).

Rips has an ingenious reply to objections of this sort. Why not, he suggests, construct a theory of belief revision that is formulated as a production system? Production systems are made up of a large number of conditional rules with specific contents. They take the form: If condition X holds, then carry out action Y, and a production can be triggered whenever its antecedent is satisfied. But, says Rips, this method of applying the rules is nearly identical to the use of *modus ponens*. Hence, the rules for belief revision and induction do obey formal rules of inference. In short, Rips proposes to promote formal rules from principles governing deduction into the fundamental principles of cognitive architecture.

The theory of recursive functions shows that a small number of different functions and a small number of different ways of combining them are sufficient to compute anything that is Turning-machine computable. Rips is proposing an analogous step for human cognition: A system of formal rules of inference specifies the "general operating principle" of the mind. What the mind does depends on how these principles are used to "program" thinking. The idea is feasible. It provides a basis for a unified theory of cognition that is an alternative, say, to Newell's (1990) SOAR theory, which is based on a production system.

In fact, Rips makes few comparisons between the Deduction-System hypothesis and other proposals about cognitive architecture. But he does discuss Newell's framework and suggests that it may suffer from two problems: It may fail to explain distinctions that are needed in accounting for inference, and "the problem-space notion may itself be too loosely constrained to be empirically helpful" (p. 28). In particular, he argues, it cannot explain the contrast between central and peripheral

processes. Whilst he sympathizes with Newell's critique of earlier theories of reasoning – that they were isolated accounts of narrow paradigms – he counters that the same might be said about the specific problem spaces invoked by Polk and Newell (in press) to account for syllogistic reasoning. The apparatus of the problem-solving approach is finally "too unconstrained to explain what is essential about deduction" (p. 30).

The problem with the Deduction-System hypothesis can be illustrated by yet another candidate for cognitive architecture – an unrestricted transformational grammar. It too can compute anything that is Turing-machine computable (Peters and Ritchie 1973). Hence, a Rips-like linguist might propose that transformational grammar specifies the "general operating" principles of the mind and that what the mind does depends on how the transformational rules are used to "program" thinking. Clearly, the critical question is: What contribution is made by postulating formal rules of inference as the basis for cognitive architecture, as opposed, say, to transformational rules? This issue must be distinguished from the empirical predictions that are made by the particular use of the rules in "programming" thinking, because what can be programmed using formal rules of inference can also be programmed using transformational rules, or production systems, or the lambda calculus, or any other universal basis for computation. Until this question is answered, it is going to be difficult to design crucial experiments that will determine the respective merits of different approaches to cognitive architecture. So, the only safe verdict about the Deduction-System hypothesis is the old Scottish one of "not proven". Let us turn to the claim that formal rules of inference do at least govern how people reason, since Rips argues that they are also demoted to play this lower-level role.

2. Reasoning as Mental Proof in a "Natural Deduction" System

At the heart of Rips's conception of deductive reasoning is the notion of a mental proof:

> I assume that when people confront a problem that calls for deduction they attempt to solve it by generating in working memory a set of sentences linking the premises or givens of the problem to the conclusion or solution. Each link in this network embodies an inference rule. . . , which the individual recognizes as intuitively sound. (Rips, p.104.)

Such proofs are analogous to proofs in formal logic, and so the task for the theorist is to devise psychologically plausible rules of inference and a psychologically plausible mechanism to use them in constructing mental proofs.

Following several proposals in the mid-1970s (e.g., Johnson-Laird 1975, Osherson 1975, Braine 1978), Rips adopts the "natural dedication" approach to rules of inference. This approach, which is due to the logicians Gentzen (1935/1969) and Jaśkowski (1934), renounces axioms in favor of rules of inference. Each logical

connective has its own rules. There are rules that introduce the connective, e.g.:

A		
B	A	A ⊢ B
∴ A and B	∴ A or B, or both	∴ If A then B

where "⊢" signifies that A leads to the derivation of B. And there are rules that eliminate the connective, e.g.:

	A or B, or both	If A then B
A and B	not-A	A
∴ B	∴ B	∴ B

Natural deduction can yield intuitive proofs, and it had a vogue in logic texts, though it seems to have been supplanted by the so-called "tree" method (e.g., Jeffrey 1981). Rips discusses the "tree" method, which simulates the search for counterexamples, but he considers it to be psychologically implausible. He writes: "The tree method is based on a reductio ad absurdum strategy" (p. 75), which he later characterizes as "unintuitive for some arguments" (p. 77). In fact, the tree method can be used to derive conclusions without the use of a reductio (see e.g., Jeffrey 1981, Ch. 2). It then appears to provide the basis for a plausible psychological theory related to the mental-model theory.

A key feature of natural deduction is the use of suppositions – sentences that are assumed for the sake of argument and that must be "discharged" sooner or later if a derivation is to yield a conclusion. One way to discharge a supposition is to incorporate it in a conditional conclusion (*conditional proof*), and another way is to show that it leads to a contradiction and must therefore be false (*reductio ad absurdum*). Thus, consider the following proof of an argument in the form known as *modus tollens*:

1. If there is a king in the hand, then there is an ace in the hand.
2. There isn't an ace in the hand.
   3. There is a king in the hand. (Supposition)
   4. There is an ace in the hand. (*Modus ponens* applied to 1 & 3)

At this point, there is a contradiction between a sentence in the domain of the premises (There isn't an ace in the hand) and a sentence in the subdomain of the supposition (There is an ace in the hand). The rule of *reductio ad absurdum* discharges the supposition by negating it:

5. There isn't a king in the hand.

Rips could have adopted a single rule for *modus tollens*, but it is a more difficult inference than *modus ponens*, and so he assumes that it depends on the chain of inferential steps illustrated here. Suppositions can be made within the subdomain of a supposition, and so on to any arbitrary depth, but each supposition must be discharged for a proof to yield a conclusion in the same domain as the premises.

The main problems in developing a natural-deduction system as a psychological theory are to ensure that it is computationally viable and that it makes sense of the empirical phenomena. An example of a computational difficulty is that the rule above for introducing "and" can run amok, leading to such futile derivations as:

A
B
∴ A and B
∴ A and (A and B)
∴ A and (A and (A and B))

and so on *ad infinitum*. Two sorts of rules are potentially dangerous: those that introduce a connective and thereby increase the length of expressions, and those that introduce suppositions. One method to prevent a rule from running amok is to do away with the rule by incorporating its effects within other rules – a method adopted for "and" and "or" introduction by the other leading formal-rule theorist, Martin Braine (see, e.g., Braine 1978). Another method for dealing with these rules is to ensure that they can be used only in preparation for the use of other major rules (Johnson-Laird 1975). A lesson from artificial intelligence, however, is that programs can use a rule in two ways: either to derive a step in a *forward* chain from its premises to its conclusion, or to derive a step in a *backward* chain from a conclusion, such as A and B, to a subgoal to prove A and a subgoal to prove B. If the program satisfies these two subgoals, then it has accomplished its main goal of proving the conjunction: A and B. The problem of controlling natural-deduction rules can be solved by using those rules that can run amok only in backward chains. Rips embodies this idea in PSYCOP, which has three sorts of rules: those that it uses forwards, those that it uses backwards, and those that it uses in either direction. The program implementing PSYCOP keeps track of both the immediate entailments from sets of sentences to their conclusion and the dependence of sentences in the derivation on earlier suppositions.

The choice of rules, Rips says, is an empirical matter. They should be ones that "the individual recognizes as intuitively sound" (p.104). Rips has canvassed previous theories, including his own, to come up with the set of rules summarized in Tables I and II. One worry about them is whether they are all intuitive. The backward rule for introducing "or", for example, was used appropriately – on Rips's own data – by only 20% of subjects. Indeed, this rule is not part of the other formal theories (e.g., Braine, Reiser, and Rumain 1984). What complicates matters is that Rips allows that individuals may differ in the rules they possess, they may learn new rules, and they may even have nonstandard rules that lead them to conclusions not sanctioned by classical logic (p.103). In an important advance over other rule theories, however, he proves two theorems about PSYCOP (based on the rules in Tables I and II). First, given an argument to evaluate, PSYCOP always halts after a finite number of steps either with a proof of the conclusion

Table I. PSYCOP'S forward rules. Certain rules, such as the one eliminating AND, are shown leading to the conclusion P; another version of the rule yields the conclusion Q. As asterisk signifies that a rule can also be used backwards.

IF P THEN Q*	IF P OR Q THEN R*	IF P AND Q THEN R
P	P	P
Q	R	Q
		R
P AND Q*	NOT (P AND Q)*	NOT (P AND Q)*
P	(NOT P) OR (NOT Q)	P
		NOT Q
P OR Q*	NOT(P OR Q)	
NOT P	NOT P	
Q		
P OR Q	NOT NOT P*	
IF P THEN R	P	
IF Q THEN R		
R		

or in a state in which it has unsuccessfully tried all available derivations. Second, PSYCOP is *incomplete* with respect to the standard sentential calculus; i.e., there are valid arguments that it cannot prove.

The strategy that PSYCOP follows in evaluating an argument with a given conclusion is to apply all its forward rules until they yield no new conclusions; it then checks whether the conclusion is among the resulting sentences. If not, it tries to work backwards from the conclusion, pursuing a chain of inference until it finds the sentences that satisfy the subgoals or until it has run out of rules to apply (p. 105). Either it succeeds in deriving the conclusion, or else it returns to an earlier choice point in the chain and tries to satisfy an alternative subgoal. If all the subgoals fail, it gives up. Here, the *incompleteness* of PSYCOP is a definite flaw. In contrasting a semantic and a syntactic method in logic, Quine (1974: 75) wrote: "[The syntactic method] is inferior in that it affords no general way of reaching a verdict of invalidity; failure to discover a proof for a schema can mean either invalidity or mere bad luck". Barwise (1993) has argued that the same problem vitiates psychological theories based on formal rules: "The 'search till you're exhausted' strategy gives one at best an educated, correct guess that something does not follow". Thus, when PSYCOP gives up, it may do so because an argument is invalid or because an argument is valid but it cannot derive the proof with its incomplete rules. The theory therefore cannot offer any account of the difference between knowing that an argument is invalid and not knowing

Table II. PSYCOP backward rules. "+" designates a supposition

+P : Q IF P THEN Q	+NOT P : Q AND (NOT Q) P	+P : Q AND (NOT Q) NOT P
P Q P AND Q	P P OR Q	
P OR Q +P : R +Q : R R	NOT (P OR Q) (NOT P) AND (NOT Q)	

whether it is valid or invalid. Yet, the remedy is at hand. As Rips remarks, PSYCOP can easily be made complete (p. 2). It calls only for the following rule:

NOT(IF P THEN Q)
P AND (NOT Q)

Rips rejects this rule because it seems paradoxical that the falsity of a conditional yields the truth of its antecedent. The founding father of formalism, Jean Piaget, had no such qualms. He wrote that if individuals have to verify whether *x* implies *y*, then they "will look in this case to see whether or not there is a counterexample *x* and non-*y*" (Beth and Piaget 1966: 181), and he invoked a formal operation of negation that had exactly the effect of the rule above. Moreover, individuals do grasp that a conditional is false when its antecedent occurs without its consequent (see Oaksford and Stenning 1992). The case for completing PSYCOP by adding this rule is strong.

A second flaw in PSYCOP is its handling of suppositions. People do make suppositions as an inferential strategy. Rips (1989) himself reports such a strategy from his earlier study of "knight and knave" problems, e.g.:

> There are only two sorts of people: knights, who always tell the truth, and knaves, who always lie. Suppose there are three individuals, A, B, and C, each of whom is either a knight or a knave. Also, suppose that two people are of the same type if they are both knights or both knaves.
>
> A says, "B is a knave."
>
> B says, "A and C are of the same type."

Question: Is C a knight or a knave?

In his description of a subject tackling this problem, Rips now writes: "This subject begins by assuming that person A is a knight. Since what A says is true on this assumption and since A says that B is a knave, the subject infers that B would be a knave" (p. 8). What is puzzling, however, is that PSYCOP does not allow this particular strategy. It places an extraordinary constraint on suppositons. As Table I shows, they can be made only when working backwards from a given conclusion, and so they cannot be used in solving knight-and-knave problems, unless a considerate experimenter presents a putative solution or the subject somehow guesses the right solution and then works backwards from it.

The constraint on suppositions causes problems in accounting for syllogistic reasoning. To explain how reasoners draw their own conclusions, Rips suggests that they formulate a tentative conclusion that allows them to work backwards to a supposition (p. 245). But this idea has not been implemented in the PSYCOP program, and it is likely to re-introduce the problems of preventing an inferential procedure from running amok – what, one wonders, is the principled distinction between supposing that a putative conclusion might be true, and supposing that a putative premise might be true? In any case, human reasoners are not constrained to making suppositions only when they already have a conclusion in mind. "Suppose everyone suddenly became dyslexic", they say to themselves, and then they follow up the consequences to an unexpected conclusion, e.g., the sale of dictionaries would decline. PSYCOP itself seems to have made a step backwards from Rips's (1989) earlier account. Unfortunately, the way to advance is not at all clear: How can suppositions be made in a forwards direction, and yet not run amok? One hint may come from theories that constrain suppositions in other ways (cf. Braine and O'Brien 1991). Another possibility is to distinguish between the *strategies* that reasoners adopt and the lower-level *mechanisms* that sanction inferential steps. One strategy is to make a supposition, but the strategic machinery must keep the lower-level mechanisms in check to prevent them losing track of the purpose of the exercise.

A third flaw arises from Rips's bias to treat deduction as a process of proving *given* conclusions. He argues correctly that experiments in which subjects draw their own conclusions may be influenced by the difficulty of putting the conclusion into words. But this methodological snag should not obscure a crucial theoretical point: Human reasoners, in contrast to most automated theorem-provers, are rather good at generating conclusions for themselves. Since infinitely many different valid conclusions follow from any set of premises, the particular conclusions that humans draw provide the investigator with important information. Their valid conclusions show that logic alone cannot account for their deductive competence, because there are many valid deductions that they never draw. They refrain, for example, from conclusions that throw semantic information away by adding disjunctive alternatives (*pace* the rule in Table 1 for the introduction of "or"). Their *invalid*

conclusions are a good clue about how the human inferential mechanism works. But Rips's bias extends to his design and analysis of experiments. He favors experiments in which the subjects are asked to evaluate given conclusions or to choose among a set of given conclusions. Strikingly, PSYCOP gives no account of what invalid conclusions occur in deduction. In discussing his own 1983 study of sentential reasoning, Rips writes:

> As a working assumption, we will suppose that errors on these problems are due to a failure to apply the rules. The failure may be due to retrieval difficulties, slips in carrying out the steps of the rule, failure to recognize the rule as applicable in the current context, or other factors. (p.153.)

Later, he even allows that reasoners' uncertainties about the correctness of a rule, or assumptions about its appropriateness, may lead to errors (p. 379). It is odd that rules that are supposed to be *intuitively sound* can get messed around in all of these ways. Nevertheless, the lesson is that errors for Rips can arise in many different ways and that systematic errors result from failures to apply rules appropriately, especially structurally more complicated rules (p. 388). As we will see, invalid conclusions turn out to expose a major defect in the theory.

3. Quantifiers in PSYCOP

Previous theories based on formal rules of inference have little to say about how people reason on the basis of quantifiers, such as "every" and "some", e.g.:

Every person has a father.
Bill Clinton is a person.
∴ Bill Clinton has a father.

The orthodox logical treatment of such inferences calls for rules that eliminate quantifiers – instantiating their variables with the names of hypothetical individuals – and for further rules that re-introduce quantifiers after inferences based on sentential connectives have been made. Rips points out that the number of possible instantiations is an exponential function of the number of variables, and so proofs soon become rapidly unwieldy. He therefore proposes to eschew an explicit representation of quantifiers, in order, he believes, to avoid some of the problems of rules that eliminate and introduce quantifiers (p. 186). It is not clear what problems Rips is avoiding, because studies of automated theorem-proving have shown that the problem of intractability re-emerges elsewhere (in the number of possible unifications). In PSYCOP's representations, the work of quantifiers is performed by names and variables in a way akin to which quantifiers are accommodated in automated theorem-provers. For example, the sentence:

Every person has a father

is represented in the following way:

IF Person(x) THEN Father(a_x, x)

where *x* stands for a universally quantified variable, and a_x stands for a temporary name with a value dependent on *x* (i.e., a so-called Skolem function). In addition to variables and temporary names, Rips allows for permanent names, such as "Bill Clinton":

Person(Bill Clinton)

He describes the elaborate transformations needed to get from a standard representation in the predicate calculus to the quantifier-free forms, but suggests that there may be a more direct route to them from sentences in natural language (p. 94). He then introduces four rules for matching expressions in his notation:

1. A variable, *x*, in a subgoal can match another, *y*, in a sentence, where both *x* and *y* derive from universally quantified variables, because logic is not affected by the particular variable representing a universal quantifier.
2. A temporary name in a subgoal can match a temporary or permanent name in a sentence.
3. A permanent name in a subgoal can match a variable in a sentence; i.e. if a sentence applies to all entities in the universe of discourse, then it applies to the particular individual in a subgoal.
4. A temporary name in a subgoal can match a variable in a sentence.

These rules are formulated with constraints to prevent invalid inferences. There is no rule, however, to match a name (permanent or temporary) in a sentence to a variable, *x*, in another sentence. Hence, the rules do not suffice to draw the conclusion:

Bill Clinton has a father

from the premises above. The problem with the required matching rule is that it too could run amok (p. 193), and so the inference can be made only by guessing the conclusion and then using rule 3 above to make the required step. Rips shows that the system is sound; that is, it does not lead to invalid conclusions. It remains, of course, incomplete in that there are valid inferences that cannot be derived within it.

Certain syllogisms are easy, e.g.:

All A are B.
All B are C.
∴ All A are C.

and:

All A are B
No B are C
∴ No A are C

And so Rips introduces two forward rules that deliver the required conclusions in a single step. He also introduces a rule of conversion:

No A are B
∴ No B are A

The similar conversion of an existential premise:

Some A are B
∴ Some B are A

is not dignified with a rule of its own, though it is intuitive. It can be derived from the rules for conjunction. The following syllogism is also easy:

Some A are B.
All B are C.
∴ Some A are C.

and even seven-year-old children can draw the conclusion for themselves. Yet, once again, the only way to proceed is to guess the conclusion and then to work backwards from it. Hence, PSYCOP would be strengthened by the introduction of more forward rules that would obviate the need to guess conclusions.

One final point about the implementation of PSYCOP: It calls for considerable sophistication. Readers can begin to grasp the problems by considering the version of *modus ponens* for constructing backward chains of inference. Here is Rips's formulation (p.197):

Backward IF Elimination (*Modus ponens*)

a. Set R to the current goal and set D to its domain.
b. If R can be matched to R′ [an isomorphic sentence that is identical apart except for variables and names] for some sentence IF P′ THEN R′ that holds in D,
c. then go to step (e).
d. Else, return failure.
e. If P′ and R′ share variables and one or more names or variables in R matched these variables,
f. then set P to the result of substituting those names or variables for the corresponding variables in P′. Label the substituting arguments, the matched arguments of P and the residual arguments the unmatched arguments of P.
g. Else, set P to P′. Label all its arguments as unmatched.
h. Apply Argument Reversal [a four-step procedure described in Rips's Table 6.4] to unmatched arguments of P.
i. Apply Subscript Adjustment [a two-step procedure described in Table 6.4] to output of Step h. Call the result P*.
j. If D does not yet contain the subgoal P* or a notational variant,
k. then add the subgoal of proving P* in D to the list of subgoals.

Who said *modus ponens* was intuitively obvious?

4. The Case for PSYCOP

Rips reviews two bodies of evidence that support the PSYCOP theory. The first set of experiments concerns sentential connectives. They include a study of a heterogeneous sample of inferences in which subjects evaluate given conclusions (Rips 1983). Rips's method is to use the PSYCOP program to find the proofs and thereby to discover which rules need to be available in order for the proofs to succeed. He then uses the experimental data to estimate the probabilities that each rule was available. The theory fits the data reasonably well. It also accounts for the times that subjects take to understand proofs laid out in explicit derivations and for their memory of proofs: They remember sentences in the same domain as the premises better than those in a subdomain based on a supposition. Rips also applies the theory to such tasks as sentence-picture verification and Wason's selection task, in which subjects have to test the truth of a conditional. With abstract conditionals, such as "If there is an A on one side of a card, then there is a 2 on the other side", they follow up the implications of a true antecedent, but not a false consequent. PSYCOP behaves similarly: It can make *modus ponens* working forward, but not *modus tollens*. With certain realistic conditionals, subjects are more likely to grasp the relevance of the false consequent. Rips makes a tentative move to invoke rules from deontic logic in order to explain the phenomenon. However, recent results by various groups of researchers have shown that the critical factor in the selection task is the availability of counterexamples (see, e.g., Green and Larking 1995; Love and Kessler 1995; and Sperber, Cara, and Girotto 1995).

The second set of experiments concerns quantifiers. Rips reports an experiment in which subjects selected one of five possible conclusions to syllogistic premises in the Scholastic figures. He uses again the method of fitting the theory to the data by estimating the probabilities that each rule was used appropriately. He shows how the theory might account for the results of a study in which subjects drew their own conclusions (Johnson-Laird and Bara 1984), suggesting, as I remarked earlier, that subjects guess tentative conclusions in order to work backwards from them. He also reports the most original of his studies – one in which he examined the response times and errors for inferences, such as:

Janet dazzled everybody.
$\therefore$ Someone dazzled somebody.

This inference calls for subjects to grasp that "Janet" implies "someone" and "everybody" implies "somebody". It therefore depends on two different matching rules, whereas an inference such as:

Everybody dazzled everybody.
$\therefore$ Fred dazzled Mary.

depends on two uses of the same rule. PSYCOP should take less time to find a rule that it has just used, and so it correctly predicted that the first inference should be

harder than the second inference (p. 250). The same prediction will be made by any theory that assumes that two inferences of the same sort should be easier than two distinct inferences.

The book focuses on these two bodies of evidence, and so it is not a general review of the psychology of deductive reasoning. It omits a number of results that PSYCOP cannot explain. They include the following robust phenomena (for a review, see Evans, Newstead, and Byrne 1993):

1. Reasoning with conjunctions is easier than reasoning with conditionals, which in turn is easier than reasoning with disjunctions (Johnson-Laird and Byrne 1991). PSYCOP can accommodate these results, but it cannot predict them.
2. The tendency to make *modus ponens* inferences can be suppressed by the presentation of certain sorts of additional conditionals (Byrne 1989).
3. Reasoning with exclusive disjunction is easier than reasoning with inclusive disjunctions (Evans et al. 1993, pp. 143–144). PSYCOP makes no special provisions for exclusive disjunctions, which presumably call for the following logical form: A OR B, and NOT(A AND B). Hence, PSYCOP seems to predict that exclusive disjunctions should be harder than inclusive disjunctions.
4. Certain diagrams – those that make alternative possibilities more explicit – both speed up and improve reasoning with disjunctions (Bauer and Johnson-Laird 1993).
5. Erroneous conclusions in most sorts of reasoning tend to be consistent with the premises; i.e., the conclusion is possibly true rather than necessarily true (Johnson-Laird and Byrne 1991). Formal-rule theorists have been known to argue that an alternative theory makes the wrong prediction, whereas their theory makes no prediction, and so their theory is superior. The folly of this position is revealed by pushing it to its logical terminus: the best theory would make no predictions at all. A whiff of this argument occurs in an experiment that Rips reports as failing to detect a difference predicted by the model theory. He argues that although the lengths of PSYCOP's formal derivations yield the same predictions as the model theory, it is "quite possible" that a rule in the shorter derivation is harder for subjects to apply (p. 368). Which, being translated, means: no conceivable result of the experiment could refute PSYCOP.

5. The Case against PSYCOP

If you ask subjects what follows from these premises:

> All the Frenchmen in the room are wine-drinkers.
> Some of the wine-drinkers in the room are gourmets.

many of them infer:

> Some of the Frenchmen in the room are gourmets.

because it is a highly credible claim (Oakhill, Johnson-Laird, and Garnham 1989). How can you convince them of their error? According to PSYCOP, you need to walk them through all possible derivations from the premises in order to show them that no derivation leads to the conclusion. That is, you need to mimic the process that leads PSYCOP to the correct response that nothing follows about the relation between Frenchmen and gourmets. If Rips is right, there is no other way to do the job. However, an effective procedure is to borrow Aristotle's device of framing premises of the same logical form with a different semantic content :

All the Frenchmen in the room are wine-drinkers.

Some of the wine-drinkers in the room are Italians.

Notwithstanding the new unified Europe, few subjects now succumb to the corresponding invalid inference (Oakhill et al. 1989):

Some of the Frenchmen in the room are Italians.

But there is another way to show subjects the error of their ways. You present a direct counterexample to the original inference. Imagine, you say to an errant subject, that there are three Frenchmen and two Italians in the room. And here you can draw a simple diagram representing the five individuals:

f
f
f
i
i

You continue: All the Frenchmen are wine-drinkers, and the Italians are wine-drinkers, too:

f w
f w
f w
i w
i w

Some of the wine-drinkers are gourmets, but now, you say, suppose that it is the Italians who are the gourmets:

f w
f w
f w
i w g
i w g

This situation satisfies the two premises but refutes the erroneous conclusion that some of the Frenchmen are gourmets, which is merely possible, not necessary. If you, the reader, grasp the force of this counterexample, then you are not reasoning according to any principle embodied in PSYCOP. You are reasoning instead according to a fundamental principle of the mental-model theory: Counterexamples are a secure – and sometimes rapid – route to establishing the invalidity of inferences.

There is a still more severe problem for PSYCOP. If you ask subjects to describe a possible hand of cards consistent with the following description:

There is a king in the hand or else there is an ace in the hand, but not both

some subjects list as a possible hand:

king

and others list:

ace

and a few include an additional card with one or other of these possibilities (Johnson-Laird and Savary 1995). This result suggests that people do not make a fully explicit representation of an exclusive disjunction. They represent that a king can occur in the hand, but they do not couple its presence with an explicit representation of the absence of an ace; they represent that an ace can occur in the hand, but they do not couple its presence with an explicit representation of the absence of a king. They may make a note to themselves to remember that the two cards are exhaustively represented and so cannot occur together, but such "mental footnotes" are easily forgotten as the load on working memory increases. Thus, for them, the two possibilities are:

king ace

rather than the fully explicit representation:

king	not-ace
not-king	ace

(Readers familiar with the mental-model theory will recognize that I am here describing it informally.) A similar tendency to make explicit as little as possible occurs if you ask subjects to describe a possible hand of cards consistent with the following conditional:

If there is a king in the hand, then there is an ace in the hand.

The preponderant response is:

king ace

Only a few subjects respond:

ace

and they do not couple it with an explicit representation that the king is not in the hand.

Such representations yield a surprising prediction. There should be "illusory" inferences with conclusions that seem compelling but that are, in fact, gross errors. For example, given the premises:

> If there is a king in the hand, then there is an ace, or else if there isn't a king in the hand, then there is an ace.

subjects should represent only the two positive cases of the disjunction, i.e., the explicit cases where the two conditionals are true:

king	ace
not-king	ace

They will therefore infer that there must be an ace in the hand. The inference is indeed compelling (Johnson-Laird and Savary 1995). Yet, it is wrong. An exclusive disjunction of the two assertions means that one of them is true and one of them is false – and the latter information, as we saw earlier, is precisely what individuals tend *not* to represent explicitly. If the first conditional is false, there is a king and no ace; and if the second conditional is false, there is no king and no ace. Either way, there is no ace. So, the subjects infer that there is an ace in the hand, when in fact it is impossible for this card to be in the hand.

This phenomenon poses a formidable problem for PSYCOP. Its rules of inference are sound, and so it cannot predict the systematic error that subjects make. One obvious maneuver is to argue that subjects misinterpret the premise – they treat the disjunction as inclusive, or they treat it as a conjunction. But an inclusive disjunction of the two conditionals (or bi-conditionals) yields a tautology. A conjunction would validly yield the conclusion that there is an ace, but this *ad hoc* assumption makes the wrong prediction that subjects will treat as contradictory the following control problem:

> There is a king in the hand and there is not an ace, or else there is not an ace in the hand and there is a king.

Moreover, illusory inferences are not limited to the earlier example; they exist in a variety of forms based on different sentential connectives.

6. Conclusions

Certain deductions seem so obvious, so automatic, so universal, that it is difficult to resist the impression that individuals have formal rules of inference for them. *Modus ponens* is the paradigm case. This intuition is the foundation of theories of reasoning based on formal rules of inference (see, e.g., Osherson 1995, Cherniak 1986, Macnamara 1986, Sperber and Wilson 1986, Pollock 1989, and Braine and O'Brien 1991). PSYCOP transcends these other theories in at least three ways. It is the first to bring together formal rules for sentential connectives and for quantifiers. It is the first to be proved to be sound. It is the first to be implemented in a computer program. Rips should be congratulated on a major achievement. He has formulated the best available formal-rule theory.

What, if anything, is wrong with PSYCOP? There are three major problems of increasing severity and generality:

1. *PSYCOP's rules are incomplete, and so it is unable to distinguish between invalidity and failure to prove validity*. It follows that people should never be able to tell that an inference is invalid. It is simple to fix this problem, because Rips describes the rule necessary for a complete logic. Even a complete theory, however, provides a suspect explanation of how reasoners decide that a conclusion is invalid. In many cases, the decision is likely to be rapid and to depend on the construction of a counterexample – a stratagem that is beyond PSYCOP's competence.

2. *PSYCOP gives an inadequate account of the conclusions that reasoners draw for themselves*. It limits suppositions to cases where there is a conclusion to be evaluated, and so it is forced to invoke an *ad hoc* guessing strategy in order to explain some simple deductions. It makes no substantive predictions about erroneous conclusions, which typically correspond to possible states of affairs rather than to logically necessary ones. Just such errors are to be expected if individuals reason, not by following formal rules of inference, but by trying to construct models of situations that satisfy premises: They overlook a model, and so their conclusion describes what is possible rather than what is necessary.

3. *PSYCOP cannot account for illusory inferences in which subjects systematically infer invalid conclusions*. Consider the epigraph of this review. Doesn't it suggest that whether Rips is right or wrong, there are rules of inference in the mind? The illusion is compelling, but in fact the assertion implies that there are no formal rules of inference in the mind. PSYCOP, however, contains only logically impeccable rules, and the only way in which they could yield invalid conclusions is by a mistake in their application. Such mistakes, as Rips rightly points out, should have "diverse sources", and so "a unified account of errors seems extremely unlikely" (p. 385). It strains credulity to imagine that errors of this sort could lead most reasoners to one and the same invalid conclusion.

Many psychologists share Rips's intuition that formal rules of inference underlie reasoning. But suppose – just suppose – that the intuition is mistaken and that the untutored mind contains no such rules. How ever could we find out? When subjects draw valid conclusions, no result is likely to undermine the intuition. Theorists can juggle the lengths of derivations and the *post hoc* availability of rules to accommodate all but the most bizarre results. If need be, they can modify the rules or the mechanism that deploys them. And, as a last resort, they can argue that the experiment did not call for the subjects to reason or that it placed a strong task demand on them to respond as if they were trying to use models or images. (One finds all these ploys in the literature.) When subjects err and make invalid inferences, an experiment may rule out theories, such as PSYCOP, that are based on sound rules of inference. Even these results, however, fail as a general case against formal rules of inference. Rule theorists could well follow Jackendoff (1988) and invoke unsound rules that deliver invalid conclusions. Rips clearly countenances the possibility: "If people possess . . . normatively inappropriate rules for reasoning with uncertainty, it seems a short step to assuming that they have similarly inappropriate rules for reasoning deductively" (p. 383). It seems an equally short step to making the theory irrefutable.

In summary, *The Psychology of Proof* is an ambitious defence of formal rules of inference, both as a basis for cognitive architecture (the Deduction-System hypothesis) and as an account of human deductive reasoning. Unfortunately, the Deduction-System hypothesis seems to have no empirical consequences that distinguish it from other claims about cognitive architecture, and so it remains no more than an interesting conjecture. The case for the theory of reasoning is more persuasive. PSYCOP makes no surprising predictions, and it has not yet led to the discovery of any striking phenomena. Its successes are more modest: It makes sense of a respectable body of data. If the reader is committed to the intuition that human inferential ability depends on formal rules of inference, then PSYCOP is unlikely to be surpassed as theory of this sort. Yet, as I have argued, it gives too weak an account of how reasoners decide that a conclusion is invalid, of how they go wrong in drawing their own conclusions, and of how they succumb to illusory inferences. For readers who see the force of these counterexamples, *The Psychology of Proof* will be a long, thought-provoking, but ultimately cautionary, tale. It cannot account for its own demise.[1]

Note

1. This work was carried out with support from the James S. McDonnell Foundation. My thanks for their many suggestions to Sam Glucksberg, Fabien Savary, and the other participants of Psychology 590: Patricia Barres, Victoria Bell, Laura Schulz, Lisa Torreano, and Isabelle Vadeboncoeur. Thanks also to Ruth Byrne for collaborating on the model theory and to Lance Rips for some helpful correctives.

References

Barwise, Jon (1993), 'Everyday Reasoning and Logical Inference', *Behavioral and Brain Sciences* 16, pp. 337–338.

Bauer, Malcolm I., and Johnson-Laird, P. N. (1993), 'How Diagrams Can Improve Reasoning', *Psychological Science* 4, pp. 372–378.

Beth, Evert W., and Piaget, Jean (1966), *Mathematical Epistemology and Psychology*, Dordrecht: Reidel.

Braine, Martin D. S. (1978), 'On the Relation between the Natural Logic of Reasoning and Standard Logic', *Psychological Review* 85, pp. 1–21.

Braine, Martin D. S., and O'Brien, David P. (1991), 'A Theory of If: A Lexical Entry, Reasoning Program, and Pragmatic Principles', *Psychological Review* 98, pp.182-203.

Braine, Martin D. S.; Reiser, Brian J.; and Rumain, B. (1984), 'Some Empirical Justification for a Theory of Natural Propositional Logic,' in Gordon H. Bower, ed., *The Psychology of Learning and Motivation*, Vol. 18, New York: Academic Press, pp. 313–371.

Byrne, Ruth M. J. (1989), 'Suppressing Valid Inferences with Conditionals', *Cognition* 31, pp. 61–83.

Cherniak, Christopher (1986), *Minimal Rationality*, Cambridge, MA: MIT Press.

Evans, Jonathan St. B. T.; Newstead, Stephen E.; and Byrne, Ruth M. J. (1993), *Human Reasoning: The Psychology of Deduction*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Gentzen, G. (1935/1969), *Investigations into Logical Deduction*, in M. E. Szabo, ed. and trans., *The Collected Papers of Gerhard Gentzen*, Amsterdam: North-Holland.

Green, David W., and Larking, R. (1995), 'The Locus of Facilitation in the Abstract Selection Task', *Thinking and Reasoning* 1:183–199.

Harman, Gilbert (1986), *Change in View: Principles of Reasoning*, Cambridge, MA: MIT Press.

Jackendoff, Ray (1988), 'Exploring the Form of Information in the Dynamic Unconscious', in M. J. Horowitz, ed., *Psychodynamics and Cognition*, Chicago: University of Chicago Press.

Jaśkowski, S. (1934), 'On the Rules of Suppositions in Formal Logic', *Studia Logica* l, pp. 5–32.

Jeffrey, Richard (1981), *Formal Logic: Its Scope and Limits; 2nd ed.*, New York: McGraw-Hill.

Johnson-Laird, P. N. (1975), 'Models of Deduction', in Falmagne, R.J., ed., *Reasoning: Representation and Process in Children and Adults*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp.7–54.

Johnson-Laird, P. N. (1983), *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*, Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N., and Bara, Bruno (1984), 'Syllogistic Inference', *Cognition* 16, pp. 1–61.

Johnson-Laird, P. N., and Byrne, Ruth M. J. (1991), *Deduction*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P. N., and Savary, Fabien (1995), 'How to Make the Impossible Seem Probable', in *Proceedings of the 17th Annual Conference of the Cognitive Science Society, Pittsburgh, PA*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 381–384.

Love, Roberta E., and Kessler, Claudius M. (1995), 'Focusing in Wason's Selection Task: Content and Instruction Effects', *Thinking and Reasoning*, 1: 153–182.

Macnamara, John (1986), *A Border Dispute: The Place of Logic in Psychology*, Cambridge, MA: MIT Press.

Marcus, S. L. (1982), 'Recall of Logical Argument Lines', *Journal of Verbal Learning and Verbal Behavior* 21, pp. 549–562.

Newell, Allen (1990), *Unified Theories of Cognition*, Cambridge. MA: Harvard University Press.

Oakhill, Jane V.; Johnson-Laird, P. N.; and Garnham, Alan (1989), 'Believability and Syllogistic Reasoning', *Cognition* 31, pp. 117–140.

Oaksford, Michael, and Stenning, Keith (1992), 'Reasoning with Conditionals Containing Negated Constituents', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18, pp. 835–854.

Osherson, Daniel (1975), 'Logic and Models of Logical Thinking', in Falmagne, R.J., ed., *Reasoning: Representation and Process in Children and Adults*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 81–91.

Osherson, Daniel N.; Smith, Edward E.; and Shafir, Eldar B. (1986), 'Some Origins of Belief', *Cognition* 24, pp. 197–224.

Peters, P. Stanley, and Ritchie, R. W. (1973), 'On the Generative Power of Transformational Grammars', *Information Sciences* 6, pp. 49–83.
Polk, Thad, and Newell, Allen (in press), 'Human Syllogistic Reasoning', *Psychological Review*
Pollock, John (1989), *How to Build a Person: A Prolegomenon*, Cambridge, MA: MIT Press.
Quine, Willard Van Orman (1974), *Methods of Logic; 3rd edition*, London: Routledge.
Rips, Lance J. (1983), 'Cognitive Processe' in Propositional Reasoning', *Psychological Review* 90, pp. 38–71.
Rips, Lance J. (1989), 'The Psychology of Knights and Knaves', *Cognition* 31, pp. 85–116.
Sperber, Daniel; Cara, Francesco; and Girotto, Vittorio (1995), 'Relevance Theory Explains the Selection Task', *Cognition*, 57: 31–95.
Sperber, Dan, and Wilson, Deirdre (1986), *Relevance: Communication and Cognition*, Oxford: Basil Blackwell.