Goals for a Theory of Deduction: Reply to Johnson-Laird

Lance J. Rips*
Department of Psychology
Northwestern University
Evanston, IL 60208
rips@nwu.edu

I wrote the book under review, *Psychology of Proof* (Rips 1994), with several goals in mind. One was to set out a theory of human deductive reasoning that had approximately the scope of first-order logic – that described inferences that depend on both sentence connectives and quantified variables. A second goal was to implement the theory as a computer program (called PSYCOP, short for Psychology of Proof) that would allow me to simulate the theory's claims about the mental steps people follow in drawing inferences in this domain. A third was to apply the theory to existing data on deductive reasoning and to test some new predictions. A fourth was to specify the theory in enough detail that it would be possible to evaluate some of its formal properties. A fifth goal was to show that the theory could also serve as the basis of a cognitive processing system, similar to production systems (e.g., Newell 1990), that could direct other forms of thought. The idea was to demonstrate in this way that deduction might be an important cognitive skill, not limited to proving mathematical theorems or solving logic brain-teasers, and perhaps show why certain deductive arguments appear to us so invincible.

These were (and are) a difficult set of goals, partly because they pull in different directions. Thanks to a century of research in logic, it isn't hard to set down a first-order deduction theory with a nice set of formal properties. And thanks to decades of research on theorem proving in artificial intelligence, we also know a little about how to streamline a first-order theory for computational purposes. But human inferences aren't streamlined, or, at least, not streamlined in the same way as contemporary theorem provers, and this means that the need to account for empirical facts about human inference can make the formal and computational properties of the theory unwieldy. To make matters worse, there were no deduction systems around (other than the formal and computational systems just mentioned) that provided guidance on how to accomplish these goals simultaneously. In the psychology of reasoning, there were proposals about how people draw inferences within some fragment of first-order logic, such as sentential logic or syllogistic logic. None of these proposals, however, came close to a full first-order theory, and few of them provided enough detail to permit rigorous evaluation. (For grumblings about the lack of explicitness in some psychological theories, see, e.g., Rips 1986, Hodges

* Thanks to Jeremy Bailenson, Norman Eliaser, and Philip Johnson-Laird for comments on an earlier draft of this response.

1993, and Bonatti 1994, as well as Chs. 9 and 10 of the book under review.)

Philip N. Johnson-Laird's review of my book in this issue touches on most of the original goals, and it provides an opportunity to reflect on how close the theory comes to fulfilling them (Johnson-Laird 1997a). In what follows, I take up Johnson-Laird's main points (though not in their original order), organizing this reply around the five aims just mentioned. Since I hope other cognitive scientists will find these goals (if not my specific solutions) to be reasonable ones, the discussion may help them see what still needs to be accomplished to obtain a worthwhile cognitive theory.

1 Goal 1: Construct a First-Order Deduction System

PSYCOP is a natural-deduction system that includes constraints on inference rules to make it psychologically realistic. One sort of constraint applies to rules like AND-Introduction ($P,\ Q \vdash P\ AND\ Q$) to prevent these rules from applying repetitively to their own output, populating memory with irrelevant information (e.g., $P\ AND\ (P\ AND\ \ldots\ (P\ AND\ Q)\ \ldots)$). PSYCOP restricts rules of this sort to situations in which the conclusion that the rule produces is a goal that the system must prove. Other rules, such as AND-Elimination ($P\ AND\ Q \vdash P$), pose no such difficulties of overproduction, and PSYCOP incorporates these rules in the traditional way: They apply whenever their premises appear in a derivation. The system uses these *forward* rules to draw inferences when it has no specific conclusion to evaluate. It uses both the forward and the *backward* (i.e., goal-restricted) rules when it must check whether a conclusion follows from a set of premises.[1]

A second difference between PSYCOP and other natural-deduction systems is that it includes no rules for quantifier introduction or quantifier elimination. The system assumes first-order representations in a Skolem-like format, using different types of symbols to capture universally- and existentially-quantified variables and their scope relations. This format is similar to that of elementary algebra, where an equation like $y = 9x-k+2$ means that there is *some* k such that for *any* x, $9x-k+2$ is y. This format appears at least as reasonable, from a psychological point of view, as the standard quantifier notation. Moreover, this representation responds to psychologists' complaints (e.g., Braine and Rumain 1983) that people don't generally follow the procedure that traditional natural-deduction systems dictate: first eliminating quantifiers in the premises (via quantifier-elimination rules), drawing propositional inferences with the resulting sentences, and then reintroducing quantifiers (via quantifier-introduction rules). These complaints motivated the change in format, not the complexity of unification or instantiation that Johnson-Laird mentions. In PSYCOP, there is no quantifier elimination or introduction because there are no quantifiers.

Of course, getting rid of formal quantifiers doesn't get rid of the need to explain inferences with universal and existential variables. We still need to explain, for

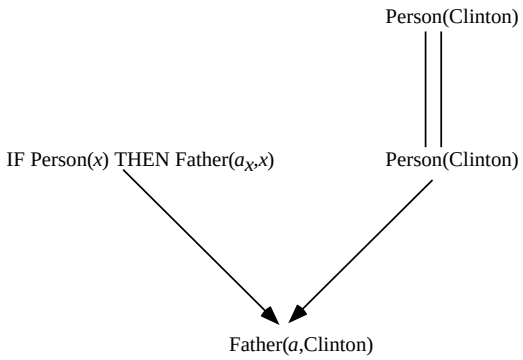

Fig. 1. PSYCOP's proof of the argument:
IF Person(x) THEN Father(a_x, x)
Person(Clinton)
Father(a,Clinton).
That is, *Every person has a father* and *Clinton is a person* entails *Clinton has a father*.

example, how we go from *Every person has a father* and *Bill Clinton is a person* to *Bill Clinton has a father*. In the new representation, this becomes:

(1) IF Person(x) THEN Father(a_x, x)
Person(Clinton)
Father(a, Clinton)

The x is a universally quantified variable; the a_x is an existentially quantified variable within the scope of x. PSYCOP handles inferences of this sort in terms of its rules for connectives – backward modus ponens in the case of the Clinton example – together with rules for matching the different types of variables and constants. Figure 1 shows the basic steps. PSYCOP matches the conclusion to the consequent of the conditional sentence, substituting *Clinton* for x. This leaves the antecedent *Clinton is a person* (i.e., *Person*(*Clinton*)). Backward modus ponens notes, at this point, that the antecedent matches one of the premises, and the conclusion therefore follows (the matching step is shown with double lines in the figure).

This method of dealing with variables, however, interacts with the forward-backward distinction we just discussed. In a system for propositional reasoning, modus ponens (*IF P THEN Q*, $P \vdash Q$) can operate harmlessly as a forward rule. Once we've made a modus ponens inference with a specific conditional sentence, we don't have to do it again, even if we've derived the antecedent in a new way. In a system for first-order reasoning, however, forward modus ponens can lead to infinitely repeated inferences. Suppose, for example, that the conditional in (1) were changed to *IF Person*(x) *THEN* (*Father*(a_x, x) *AND Person*(a_x)) – every person has a father who's a person. Then, from the fact that Clinton is a person, we can derive by forward modus ponens that his father is a person. By a second

application of modus ponens to the same conditional, his father's father is a person; by a third, that his father's father's father is a person. And so on, up the family tree.

This difficulty with forward inferences in a context with variables presents a dilemma for cognitive theory. We clearly can't allow the rules to fill memory with an unending stream of irrelevant information. So either we have to eliminate the forward version of the rule, retaining the backward version to restrict the number of inferences, as we did with AND-Introduction, or we can keep the forward rule and place some external constraints on it, for example, limiting the form of the conditionals to which it can apply or the number of times it can apply to a single conditional.[2] PSYCOP handles this dilemma by compromise, retaining forward modus ponens when there are no variables to bind, but restricting modus ponens to its backward form when it must bind variables in the antecedent of the conditional (e.g., *IF* $P(x)$ *THEN* $Q(x)$) to those of the minor premise (e.g., $P(a)$). The theory also allows the possibility that certain contexts will suggest the form that a conclusion can take, with this form then triggering the backward rule to confirm or disconfirm the suggestion. In reasoning tests using Aristotelian syllogisms, for example, there are only a small number of possible conclusions, and PSYCOP can check these. Indeed, it's hard to imagine how participants in such an experiment could *keep* from hypothesizing an answer in these conditions, and the mistakes the participants make are often interpretable in this way (as shown in Ch. 7 of the book).[3]

This compromise solution overcomes the problem of infinite irrelevant conclusions, but it also means that PSYCOP will not *automatically* draw the conclusion of (1), given just its premises. (PSYCOP will, of course, affirm that the conclusion follows from the premises if it inspects the entire argument.) I know of no data that suggest that this solution is incorrect, but, like most compromises, this one may require tinkering, perhaps by adding more forward rules (as Johnson-Laird suggests) or by limiting how PSYCOP applies its rules (as Braine, O'Brien, et al. 1995 suggests). This is one place where more refined experiments will help in understanding how people cope with restrictions on their inference abilities.

2 Goal 2: Implement the Theory as a Computer Program

PSYCOP is embodied in a computer program that draws inferences using the procedures just sketched. The program is itself "programmable", since it is possible to write procedures in the PSYCOP language for cognitive tasks such as classification or problem-solving. Chapter 8 of *Psychology of Proof* ("The Role of Deduction in Thought") gives PSYCOP programs for solving John R. Hayes's (1965) spy problems and for carrying out simple classification. In more recent work (Rips 1995), I give a PSYCOP program for the Tower of Hanoi problem.

A computer simulation of the theory enables a more detailed account of the individual deduction steps than usually appears in cognitive papers. This seems

beneficial, on the whole, since it provides a more accurate view of how the theory works and will enable others to give more informed evaluations and improvements. One disadvantage of this level of detail, however, is that it may give the illusion that, according to the theory, deduction is ordinarily a sophisticated, unintuitive, conscious process. But detail in the description needn't imply that the routines are hard for people to use, any more than a detailed description of the visual system implies that objects are hard for people to perceive. The ability of modus ponens, for example, to bind variables in the current theory means that this inference requires several steps for arbitrarily complex conditional sentences (which may have variables local to the antecedent, variables local to the consequent, and shared variables in complex scope relations). How complex these steps are in practice, however, depends on the sentences to which they apply and on the cognitive resources they consume. Johnson-Laird cites the full version of backward modus ponens that is needed for arbitrary sentences. For simple conditionals, however, PSYCOP can omit some of the steps entirely, and some of the remaining steps also simplify, as in the Figure 1 example. There is no reason to suppose that these procedures are unintuitive or out of reach for most cognizers. (Of course, the technical description of a cognitive process is not necessarily a good guide to its intuitive complexity. To readers who aren't familiar with the technical vocabulary, the description will seem complicated, even if the underlying cognitive steps are simple in themselves.)

PSYCOP executes rules like backward modus ponens both on a nonconscious architectural level and on a conscious level in deliberate reasoning. Johnson-Laird (personal communication, 1996) raises the issue of whether nonconscious procedures can be as sophisticated as conscious ones, but it is hard to see how cognitive psychology could make much progress if it were to limit nonconscious information-processing to simple routines. Surely, motor control, perception, sentence recognition, sentence production, categorization, recognition memory, and many other cognitive abilities depend on nonconscious processes of formidable complexity, and it would be astonishing if reasoning were an exception to this trend.

3 Goal 3: Apply the Theory to Empirical Data

The deduction theory is applied to a wide range of deductive arguments, including simple propositional arguments, classical syllogisms, arguments that depend on universal and existential variables, and arguments that depend on the full first-order mix of variables and connectives. Many of these experiments present the arguments to groups of people and tabulate the percentage of times they decide that the conclusion follows logically from the premises. In general, the model predicts the likelihood of a correct answer in terms of the likelihood of correctly applying the forward or backward rules that PSYCOP needs in its derivation. The greater the number of rule applications in the proof (and the smaller the likelihood of applying each rule correctly), the smaller the likelihood that people

will correctly appraise the argument.

These assumptions fit the individual data sets quite well, but what's most interesting about the results in the book is the stability *across* experiments. The estimates of how likely it is that people will correctly apply a rule are in good agreement from one experiment to the next, despite changes in participants, wording conventions, and problem type (e.g., from propositional-logic problems to syllogisms to predicate-logic problems). The estimates for different rules also accord well with our intuitive notions of the difficulties of these rules, with AND-Introduction and -Elimination being consistently easier than IF-Introduction (conditional proof), and rules for NOT-Introduction (proof by contradiction) being consistently among the most difficult. This evidence converges nicely in supporting the claim that people apply mental proof-rules in evaluating arguments.[4]

In many of these experiments, participants see complete arguments and judge whether the conclusions of these arguments follow. Other experiments in the book, however, employ memory tasks, proof-following tasks, and liar–truth-teller puzzles. The book also applies the model to an earlier experiment of Johnson-Laird and Bruno Bara (1984) in which the participants supplied their own conclusions for the premises of categorical syllogisms. As Johnson-Laird notes, however, the book favors experiments in which participants evaluate arguments rather than produce conclusions. One reason for this, as Johnson-Laird concedes, is that production data have the "methodological snag" that they are influenced by difficulties people have in framing the conclusions in natural language. (See Greene 1992 for a potential case of this sort, and Johnson-Laird, Byrne, and Tabossi 1992 for a reply.) A deeper reason is that production experiments aren't very sensitive tools in determining the limits of people's deduction abilities (as Bonatti 1994 argues). There's little motivation for participants to formulate more complicated conclusions when simpler ones satisfy the requirements of the task, and the conclusions they do produce are subject to response bias from the form of the premises, from conversational demands, and from task demands. Of course, this doesn't imply that there is nothing to be learned from such experiments (that's why I fit the Johnson-Laird and Bara data) nor that evaluation experiments are methodologically pure (since no experiments are).

Johnson-Laird's central complaint about PSYCOP is closely connected with his focus on production data. He believes PSYCOP "gives no account of what invalid conclusions occur in deduction", "cannot predict the systematic error that subjects make", "gives an inadequate account of the conclusions that reasoners draw for themselves", and "cannot account for illusory inferences in which subjects systematically infer invalid conclusions" (Johnson-Laird 1997a). These last two points are two of the three "major problems" that he cites against the theory (the third is considered in the following section) and the ones he considers most severe and general. As Johnson-Laird correctly states, the theory identifies many sources of errors in deduction. Indeed, the history of psychological research in this field, for better or worse, is largely the story of uncovering these sources, and we now

know a large number of them: incorrect interpretation of instructions, incorrect interpretation of premises due to people's prior knowledge of the subject matter, production errors due to difficulty in "putting the conclusion into words", errors due to working-memory limitations, bias due to prior belief in the conclusion, conversational implicatures, and many more. It is possible to debate whether these "errors" impugn the rationality of people who make them (see Ch. 11 for a discussion), and it is possible to debate which errors account for which responses. But it would be difficult to maintain that these varied sources don't contribute, at least on some occasions, to the answers people give in reasoning experiments – a point on which Johnson-Laird and I agree. Since these errors have obviously disparate causes, there can't be a unified theory of deduction errors, any more than there can be a unified theory of automotive breakdowns. Sometimes it's the ignition system; sometimes it's a flat tire; sometimes it's an engine rod. There is no natural kind consisting of all and only automotive "errors", and, likewise, no natural kind consisting of all and only reasoning errors (i.e., errors committed in reasoning experiments).

Is there a narrower domain of reasoning errors or reasoning difficulties that a scientific theory of deduction must explain? Johnson-Laird (1997a) mentions a number of findings that he believes create difficulties for PSYCOP. Although he discusses these as if they were well-documented results, a closer look at these effects suggests that they are quite controversial, sometimes directly contradicting each other. It's worth reviewing this evidence on a point-by-point basis:

Reasoning with connectives: Johnson-Laird asserts that "[r]easoning with conjunctions is easier than reasoning with conditionals, which in turn is easier than reasoning with disjunctions". However, the data don't support any such generalization. One can, of course, find arguments whose difficulty accords with this ordering, but it is equally possible to find arguments that contradict it. For example, subjects rate as more difficult the argument *NOT*(*P AND Q*); *P*; *Therefore, NOT Q* than the argument *P OR Q; NOT P; Therefore, Q* (2.35 vs. 3.10 on a 9-point scale of rated difficulty in Braine, Reiser, and Rumain 1984). Likewise, arguments with the same connective can differ drastically in difficulty. This fact is easy to accommodate in theories like PSYCOP, since they handle different entailments with different rules. (For a further counterexample, see *Psychology of Proof*, Ch. 10.)

Reasoning with exclusive versus inclusive disjunction: According to Johnson-Laird, reasoning with exclusive *or* (*XOR*) is easier than reasoning with inclusive *or* (*OR*), and he cites Jonathan St. B. T. Evans et al.'s (1993) review as support for this conclusion. However, a look at Evans et al.'s tabulation of results shows no difference across studies. Their Table 5.4 lists the argument *P OR Q; NOT P; Therefore, Q* as correctly evaluated by 80% of subjects, while *P XOR Q; NOT P; Therefore, Q* is correctly evaluated by 84% of subjects. (Medians are 83% and 84%.) Of course, many more subjects endorse *P XOR Q; P; Therefore, NOT Q* than *P OR Q; P; Therefore, NOT Q* (Evans et al. 1993, Table 5.5), but

that's because only the former is valid.

Representation of exclusive disjunction: Johnson-Laird believes that "people do not make a fully explicit representation of an exclusive disjunction." As evidence, he offers the fact that

> 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 . . . .

What the subjects don't do is explicitly mention the fact that there is no ace in the first hand above and no king in the second. But surely this is due to a strong conversational demand to list the cards *in* the hand rather than the cards *not* in it. No inference can be drawn about people's mental representations from such data, a further example of the difficulties inherent in production experiments.

Suppression of modus ponens: Ruth Byrne (1989) found that people sometimes withhold a modus-ponens inference (e.g., *If it's sunny, we'll go to the beach; It's sunny; Therefore, we'll go to the beach*) if they read certain additional premises (e.g., *If I remember my beach pass, we'll go to the beach*). However, the cause of this finding is hotly debated (see Politzer and Braine 1991), and it would be premature to give a theoretical account before further experiments resolve this issue.

Effects of diagrams: Diagrams can aid reasoners in proper settings. External spatial representations impose constraints that can be helpful when they coincide with problem constraints, and they may aid cognitive bookkeeping by organizing problem information. It is unclear, however, how well training with diagrams transfers to other types of reasoning problems (see Stenning et al. 1995). It's also controversial whether people spontaneously (i.e., without special training) use mental diagrams in reasoning (see Ch. 10 for a review). Recent findings suggest that not all subjects produce spatial diagrams in solving syllogisms; for those who do, the diagrams tend to be variations on the Euler circles that they've learned in school (Ford 1995; see also Sect. 5, below).

Consistent conclusions and illusory inferences: Two final effects deserve special treatment, since they highlight the difficulties in constructing deduction theories. According to Johnson-Laird, (a) deduction theories must be able to explain why the invalid conclusions that people draw on their own are logically consistent with the premises, and (b) they must account for "illusory inferences

in which subjects systematically infer invalid conclusions", such as the following example, which he draws from Johnson-Laird and Savary 1995:

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

There is an ace.

From the premises of (2), people tend to draw the conclusion shown above, even though this conclusion is invalid.

A grave difficulty for Johnson-Laird's brief arises at this point, however. The error tendencies in (a) and (b) are themselves contradictory and therefore can't be universally true. As Johnson-Laird clearly shows, the conclusion of Argument (2) is logically *inconsistent* with its premises, and it thus provides a straightforward counterexample to the alleged tendency in (a) to make errors that are consistent with the premises. No consistent theory could possibly account for both (a) and (b) as invariable tendencies.

Indeed, one of the most remarkable aspects of Johnson-Laird's clever new finding is that, whatever the problems that "illusory inferences", such as (2), pose for other theories, they devastate the theory presented in Johnson-Laird, Byrne, and Schaeken 1994. In that paper, Johnson-Laird et al. have this to say about the sort of data that would refute mental-model theory:

> What would contravene the fundamental principles of the model theory? In fact, ... the theory is simple to refute in principle. It applies to all domains of deduction, and it makes two general predictions about them. The first prediction is easy to test, because it does not call for any account of the specific models for a domain. According to this prediction, erroneous conclusions should be consistent with the premises rather than inconsistent with them, because reasoners will err by basing their conclusions on only some of the models of the premises. They will accordingly draw a conclusion that is possibly true rather than necessarily true .... (p. 735.)

To the extent that subjects endorse the conclusion of (2), as Johnson-Laird assures us they do, they thereby "contravene the fundamental principles of model theory".

This doesn't mean that people *never* exhibit the types of errors mentioned in (a) and (b). In line with (a), PSYCOP would expect a greater number of errors that are consistent rather than inconsistent with the premises, because it can check and eliminate certain types of inconsistencies. As mentioned earlier, errors can arise in many ways, according to the theory, but, for these purposes, let's distinguish errors that stem from people's initial misunderstanding of the premises and those that stem from later parts of the deductive process – for example, priming of conclusions by the premises or misapplication of logical rules. If these latter processes yield an erroneous *inconsistent* conclusion, PSYCOP can apply its backward rules to the negation of this conclusion to eliminate it. A proof of the negation of the conclusion demonstrates that the original conclusion is inconsistent. No such strategy will work with an erroneous *consistent* conclusion, however: There is no possible proof

of the negation of such a statement, and PSYCOP could not distinguish it on this basis from a correct inference. How often people engage in this type of consistency checking probably depends on the demands of the task. Conditions that place heavy emphasis on accuracy and that allow unlimited time per problem should increase the amount of checking and decrease the number of inconsistent errors.

Notice, too, that this checking process is inherently unable to eliminate errors due to misinterpretation of premises. Erroneous conclusions of this sort follow validly from the (misinterpreted) premises and will therefore not be identified as inconsistent. This seems to be a potential explanation for errors on Argument (2), since the premises seems to invite the interpretation *If there is either a king or no king in the hand, then there is an ace*, from which the conclusion validly follows. This possibility is speculative, however: Johnson-Laird and Fabien Savary (1995) provide only preliminary results on such inferences.[5]

Johnson-Laird has changed his view about the consistency of conclusions because of arguments such as (2). The passage from Johnson-Laird et al. 1994, quoted above, no longer represents his current theorizing (personal communication, 1996). His present view is that some initial models support conclusions consistent with the premises, whereas others support conclusions that contradict the premises. The point of discussing the argument is not to show that Johnson-Laird's present position is contradictory, but to highlight the difficulties inherent in his review. There are many systematic sources of error, including contradictory ones such as (a) and (b). Thus, deduction theories must choose which errors to explain internally and which to explain as the effects of other cognitive processes (e.g., comprehension or response processes). There are certainly sources of systematic error that PSYCOP doesn't explain internally and, likewise, sources that Johnson-Laird's theory can't explain. The latter include, for example, difficulties people have with the scope of logical operators and difficulties associated with instantiating variables, since Johnson-Laird's mental models don't include information about either scope or variables.[6] Which error tendencies a deduction theory *should* explain is not a matter that can be decided simply by listing errors, but depends on the entire cognitive architecture in which deduction takes place.

4 Goal 4: Assess the Theory's Formal Properties

It's helpful in understanding a system as complex as this one to compare it to standard logical benchmarks. One reason for this is simply to ensure that the system is consistent. However, it is also useful to know how the system compares in logical power with other well-known logics. Can it, for example, prove all the theorems that are derivable in classical predicate logic? For PSYCOP, the answer is no, according to proofs in Chapters 4 and 6 ("Mental Proofs and Their Formal Properties" and "Variables in Reasoning", respectively): There are arguments containing material conditionals that the program can't prove. Johnson-Laird regards this as a flaw in the system, but it is difficult to see why. PSYCOP is intended as a theory

of human reasoning and not a theorem prover for logic. Given the controversies surrounding the material conditional, it would be odd if the semantics of that connective coincided with people's logical intuitions about natural-language "if", and the handling of conditionals is precisely where PSYCOP and classical logic part company. PSYCOP is complete for a classical propositional logic in which material conditionals are replaced by equivalent truth functions containing AND, OR, and NOT. (Johnson-Laird maintains that people recognize entailments of the form $NOT(IF\ P\ THEN\ Q) \vdash P$, which are not included in PSYCOP. The evidence he cites, however, supports, not the entailment just mentioned, but instead $P\ AND\ NOT\ Q \vdash NOT(IF\ P\ THEN\ Q)$, with which PSYCOP has no difficulty.)[7]

I suspect, however, that the target Johnson-Laird has in mind in criticizing PSYCOP on formal grounds is not its *incompleteness* with respect to classical logic, but its *undecidability*. The passage that he quotes from Quine 1982, p. 88, for example, is explicitly about decidability – the property a proof system has if, for any argument, it halts after a finite number of steps with a verdict about the argument's deductive correctness or incorrectness.[8] It follows as a corollary of the proofs in Chapter 4 that PSYCOP is a decision procedure for classical propositional logic formulated with AND, OR, and NOT. For richer systems, such as predicate logic, PSYCOP is, of course, not decidable. But decidability with respect to classical predicate logic is not something one can hold against PSYCOP, since it is something that no mechanistic system can attain (according to Church's Thesis and Church's Theorem). In this respect, PSYCOP is in exactly the same boat as all other computable theories of deduction, including Johnson-Laird's own theory of mental models. Moreover, in these richer systems, people often do find themselves wondering whether their inability to find a proof (on a math test, for example) means that no proof exists or that they simply haven't been able to find it.

Of course, this isn't to suggest that no more needs to be done to understand the system's formal properties. Although the book contains proofs of the soundness of some of the inference rules, we still need proofs for others. And although I show that PSYCOP doesn't prove all classically valid theorems with conditionals, it remains to be seen whether there is a natural semantics for its conditional connective.

5 Goal 5: Show How the Theory Can Serve as a Cognitive Processing System

PSYCOP hoped to make plausible the idea that deduction could serve as the basis for other cognitive processes. Production-system theories of cognition, such as Anderson 1983, Holland et al. 1986, and Newell 1990 prepared the ground for such a claim. According to these theories, long-term memory contains many thousands of conditional commands, similar to the IF-THEN statements in conventional programming languages. When the contents of a temporary working memory meet

the conditions of one or more commands, the commands carry out an operation, usually adding more information to working memory. A production system applies the commands by matching variables to working-memory symbols and then carrying out a practical form of modus ponens, executing a computational step. Because PSYCOP can bind variables and perform modus ponens, it can serve as a means of directing other mental operations in a similar way. Moreover, PSYCOP has many other logical rules as well; so it can provide a programming structure that might be more flexible and psychologically realistic than production systems.

This ability to direct other cognitive processes seems to me the right way for PSYCOP to account for skills that people sometimes learn in order to supplement their basic deduction abilities. There is no doubt, for example, that people can learn devices like Euler circles or Venn diagrams and can use them to test syllogisms by searching for counterexamples. With practice, they can learn to manipulate these diagrams mentally, just as they can learn to do mental multiplication. Similarly, in solving liar–truth-teller problems (e.g., Rips 1989), people must be able to use the problem instructions to devise a strategy to keep track of temporary assumptions and to draw special-purpose inferences from them. In this context, it is possible for PSYCOP to monitor the assumptions and inferences in order to curb their potential to consume too many working-memory resources. Contrary to Johnson-Laird's assertion that PSYCOP's handling of assumptions is a retreat from the model in Rips 1989, its new capability is much more general.[9] It has the same advantages, in this respect, that production-system models have over ad-hoc models for specific mental skills.[10]

There are, of course, many ways to compute the same input-output functions. You can do it with a Turing machine, a set of recursive functions, a production system, or even a transformational grammar, as Johnson-Laird observes. To decide what sort of system underlies cognition, you need to determine the atomic mental steps people go through in carrying out cognitive tasks. You can then select the system that makes these steps available (see, e.g., Pylyshyn 1989). This is a tall order, to say the least – tantamount to solving most of the key questions in cognitive psychology. I hope to have shown that PSYCOP offers a reasonable way to explain what people do in solving specifically deductive problems, but claims about cognitive architecture go far beyond these tasks. To provide some plausibility for the more sweeping claim, Chapter 8 of the book contains examples of how PSYCOP can solve simple search and categorization problems, as mentioned earlier. PSYCOP can also solve the Tower of Hanoi puzzle, a chestnut in the problem-solving literature, using a strategy that Herbert A. Simon (1975) attributed to human subjects (see Rips 1995). These illustrations are far from a proof that the deduction system is the right architecture, but they provide a start.

6 Summary

A big advance in the psychology of deductive reasoning was Daniel N. Osherson's (1974–1976) proposal of an algorithm to account for children's inferences. Although his series of books met mostly criticism at the time (including his own disarming self-criticism), it's clear that they pioneered unified theories that respond to both formal and empirical issues. This comprehensive approach helped avoid the lack of progress in the psychology of reasoning that had been caused by bickering over small-scale models, and they set the stage for current theorizing. PSYCOP's goals, at their most ambitious, further enlarge the scope of deduction theory. It aims to provide a theory for first-order reasoning – one that is well-specified formally and computationally, that accounts for human inference patterns, and that can serve as a basis for other cognitive tasks. Many aspects of the theory still need revisions or extensions, especially in applying the theory to new types of data, in establishing the theory's formal semantic properties, and in experimenting with its uses as a cognitive architecture. Accomplishing all this requires integrating knowledge from logic, artificial intelligence, and cognitive psychology – a daunting task, considering the amount of information available in these fields. The effort of integration is worthwhile, however, if only because it provides a clearer view of the serious issues.

Notes

[1] PSYCOP's forward-backward distinction came from analyzing the strategies students use to prove theorems in standard natural-deduction systems, and it is also consistent with similar goal-triggered versus premise-triggered methods in artificial intelligence. The distinction has a long history, however.

[2] The solution by external constraints is, roughly speaking, the one adopted by production systems. I discuss such systems later in connection with Goal 5. The solution by internal constraints is close to the one adopted by logic-programming languages, such as Prolog.

[3] Of course, this shouldn't be taken to imply that people never make mistakes on syllogisms that have valid conclusions. According to PSYCOP, whether they find these conclusions depends on at least two factors. First, people may check only those potential conclusions that the premises prime. For example, premises of the form:

No A are B.
All B are C.

may suggest *No A are C* (rather than, e.g., *Some C are not A*) as a conclusion, because the form of this tentative conclusion matches that of the first premise. Second, even when people come up with what is in fact a correct hypothesis, they must still verify that this conclusion follows, and the verification process is susceptible to error. See Ch. 7 of the book under review for a full account of syllogisms.

[4] One aspect of the model fitting that some readers have found odd is that when PSYCOP's predictions are fit, the parameter estimates for some rules are relatively low, indicating that subjects don't often apply these rules in relevant proofs. For example, although PSYCOP contains a rule for OR-Introduction ($P \vdash P\ OR\ Q$), the data indicate that subjects successfully apply this rule only about 20% of the time. Why such a low value if the rule is a basic one? One plausible answer is that in many situations the rule violates Gricean conversational principles, and this may cause subjects to avoid it. If you know that P is true, it seems conversationally inappropriate to assert only the weaker statement $P\ OR\ Q$. In other settings, however, the rule seems secure. If you have to choose between a bet that P is true and a bet that $P\ OR\ Q$ is true, you would

presumably choose *P OR Q* on the grounds that any situation in which *P* is true must also be one in which *P OR Q* is true. This application of OR-Introduction seems reasonable, since it doesn't violate Gricean maxims. For further discussion of constraints on PSYCOP's parameters, see Ch. 11 of *Psychology of Proof*.

[5]Johnson-Laird believes that the above hypothesis seems less plausible when applied to a variation of (2): *One of these assertions is true and one of them is false: If there is a king in his hand, then there is an ace in his hand. If there is not a king in his hand, then there is an ace in his hand* (see his reply to this paper, Johnson-Laird 1997b). But it is hard to see why the hypothesis should be any less plausible with the "one is true and the other is false" wording than with the original "or else". Johnson-Laird also points out that the explanation doesn't cover incorrect conclusions that people draw from the following premises: *Only one of the following assertions is true: Albert is here, or Betty is here, or both. Charlie is here, or Betty is here, or both. This assertion is definitely true: Albert isn't here and Charlie isn't here.* It's not obvious that both sorts of errors are the result of the same reasoning process. However, one alternative possibility that would cover both is that people take the disjunctive phrasing (*or else*, *one is true and one is false*, *only one is true*) as an invitation to apply OR-Elimination, and they interpret the rest of the argument opportunistically to fit this schema. The schema states that if a statement follows from *P* and it also follows from *Q*, then it follows from the disjunction *P OR Q* (see *Psychology of Proof*, Chs. 2–4). For the conditional problems, people might reason: "Either he has a king or no king. If he has a king, then he has an ace. If he has no king, then he has an ace. So, either way, he has an ace." On the second problem, the reasoning might be: "Either (Albert is here or Betty is here) or (Charlie is here or Betty is here). If the first possibility holds, then Betty is here (since Albert isn't). If the second possibility holds, then Betty is also here (since Charlie isn't). So, in either case, Betty is here." This hypothesis accords well with the line that people take when I've asked them informally to think aloud about the problems. Like the hypothesis mentioned above, however, it is tentative, since Johnson-Laird's data on these problems haven't appeared. In general, it seems an odd strategy for Johnson-Laird to rest his case for mental models on results that at present are either "in press" or "submitted".

[6]According to Johnson-Laird and Byrne's theory (1991, Ch. 9), people first translate quantified sentences in natural language, such as *All x's are equal to the sum of some y and some z*, into sentences of predicate logic (e.g., $(All\ x)(Some\ y)(Some\ z)(x = y + z)$). To form a mental model, however, people are supposed to translate this sentence once more into a specific instantiation in which the quantifiers and variables are dropped. For the above sentence, Johnson-Laird and Byrne give the mental model as:

$$x = ([8]\ [6]),\ y = (1\ 6\ 4\ 2),\ z = (7\ 7\ 2\ 2\ 4\ 4)$$

It's an important question whether this mental model is anything like an adequate representation of the original sentence. The point here, however, is that the mental models themselves don't include scope relations or variables. Hence, scope and variables play no role in the reasoning process, which is supposed to be solely a matter of manipulating mental models. If there is any remaining doubt about this, Johnson-Laird (1989, p. 488) *defines* a mental model as a representation that meets three conditions, one of which is "Unlike other proposed forms of representation, [a mental model] does not contain variables." (See Chs. 7 and 10 of *Psychology of Proof* for discussions of these problems.)

[7]This problem carries over to Johnson-Laird's reply to this article (1997b). He states: "Yet several experimenters have asked people to construct instances that would falsify a conditional of the form: if *p* then *q*. They reliably respond with cases of: *p* and not *q* (see, e.g., Oaksford and Stenning 1992)." But what matters here is not what *falsifies* a conditional, but what *follows from* a false conditional, two very different matters.

[8]Johnson-Laird's use of this quotation is misleading in a second way. Quine (1982, p. 88) is not there comparing syntactic methods in general to semantic methods, but comparing the axiomatic method to other methods (both syntactic and semantic) that yield a decision procedure. The full passage reads as follows:

> On the other hand an axiom system for logic is necessarily foundational, and I would in conclusion remark further that it is of dubious value – especially in the logic of truth functions.

> This domain, after all, enjoys the luxury of a decision procedure for validity – that is, a mechanical test. Truth-value analysis affords one such test of validity; truth tables afford another; transformation into conjunctive normal form affords a third. Thus blessed, we should be unwise to make practical use of the axiomatic method in this domain. It 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.

[9]In particular, PSYCOP retains the ability to reason both forward and backward from assumptions. The only restrictions on forward reasoning are those discussed earlier that keep the system from producing an infinite number of conclusions. In his reply to this paper, Johnson-Laird focuses his critique on the fact that people (but not PSYCOP) can make arbitrary assumptions with no goal in mind. Making unmotivated assumptions in tackling a logic or a math problem, however, is a sure way to fail to solve it. In a brain-storming session, people might consider in a relatively undirected way what would happen if everyone suddenly became dyslexic (to use Johnson-Laird's example), but it is better to handle this behavior as a separate skill rather than building it into the deduction component.

[10]In his reply to this article (1997b), Johnson-Laird complains that this generality makes the theory irrefutable. In this respect, though, it is on exactly the same footing as other proposals for cognitive architectures, such as production systems. It is quite true that PSYCOP, production systems, and other general architectures could simulate mental models if the mental-models theory were spelled out in adequate detail. This does not make these theories immune to refutation, however, for reasons that Zenon Pylyshyn (1989) has described.

References

1. Anderson, John R. (1983), *Architecture of Cognition*, Cambridge, MA: Harvard University Press.

2. Bonatti, Luca (1994), 'Propositional Reasoning by Model?', *Psychological Review* 101, pp. 725–733.

3. Braine, Martin D. S.; O'Brien, David P.; Noveck, Ira A.; Samuels, Mark C.; Lea, R. Brooke; Fisch, Shalom M.; and Yang, Yingrui (1995), 'Predicting Intermediate and Multiple Conclusions in Propositional Logic Inference Problems: Further Evidence for a Mental Logic', *Journal of Experimental Psychology: General* 124, pp. 263–292.

4. Braine, Martin D. S.; Reiser, Brian J.; and Rumain, Barbara (1984), 'Some Empirical Justification for a Theory of Natural Propositional Logic', *Psychology of Learning and Motivation* 18, pp. 313–371.

5. Braine, Martin D. S., and Rumain, B. (1983), 'Logical Reasoning', in John H. Flavell and Ellen M. Markman, eds., *Cognitive Development: Handbook of Child Psychology, Vol. 3*, New York: John Wiley, pp. 263–340.

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

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

8. Ford, Marilyn (1995), 'Two Modes of Mental Representation and Problem Solution in Syllogistic Reasoning', *Cognition* 54, pp. 1–71.

9. Greene, Steven B. (1992), 'Multiple Explanations for Multiply-Quantified Sentences: Are Multiple Models Necessary?', *Psychological Review* 99, pp. 184–187.

10. Hayes, John R. (1965), 'Problem Topology and the Solution Process', *Journal of Verbal Learning and Verbal Behavior* 4, pp. 371–379.

11. Hodges, Wilfrid (1993), 'The Logical Content of Theories of Deduction', *Behavioral and Brain Sciences* 16, pp. 353–354.

12. Holland, John H.; Holyoak, Keith J.; Nisbett, Richard E.; and Thagard, Paul R. (1986), *Induction: Processes of Inference, Learning, and Discovery* Cambridge, MA: MIT Press.

13. Johnson-Laird, P. N. (1989), 'Mental Models', in Michael I. Posner, ed., *Foundations of Cognitive Science*, Cambridge, MA: MIT Press, pp. 469–499.

14. Johnson-Laird, P. N. (1997a), 'Rules and Illusions: A Critical Study of Rips's *The Psychology of Proof*', *Minds and Machines* 00, pp. 000–000.

15. Johnson-Laird, P. N. (1997b), 'An End to the Controversy? A Reply to Rips', *Minds and Machines* 00, pp. 000—00.

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

17. Johnson-Laird, P. N.; Byrne, Ruth M. J.; and Schaeken, Walter (1992), 'Propositional Reasoning by Model', *Psychological Review* 99, pp. 418–439.

18. Johnson-Laird, P. N.; Byrne, Ruth M. J.; and Schaeken, W. (1994), 'Why Models Rather than Rules Give a Better Account of Propositional Reasoning: A Reply to Bonatti and to O'Brien, Braine, and Yang', *Psychological Review* 101, pp. 734–739.

19. Johnson-Laird, P. N.; Byrne, Ruth M. J.; and Tabossi, Patrizia (1992), 'In Defense of Reasoning: A Reply to Greene', *Psychological Review* 99, pp. 188–190.

20. 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.

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

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

23. Osherson, Daniel N. (1974–1976), *Logical Abilities in Children (Vols. 2–4)*, Hillsdale, NJ: Lawrence Erlbaum Associates.

24. Politzer, Guy, and Braine, Martin D. S. (1991), 'Responses to Inconsistent Premises Cannot Count as Suppression of Valid Inferences', *Cognition* 38, pp. 103–108.

25. Pylyshyn, Zenon (1989), 'Computing in Cognitive Science', in Michael I. Posner, ed., *Foundations of Cognitive Science*, Cambridge, MA: MIT Press, pp. 49–91.

26. Quine, Willard Van Orman (1982), *Methods of Logic, 4th edition*, Cambridge, MA: Harvard University Press.

27. Rips, Lance J. (1986), 'Mental Muddles', in Myles Brand and Robert M. Harnish, eds., *The Representation of Knowledge and Belief*, Tucson, AZ: University of Arizona Press, pp. 258–286.

28. Rips, Lance J. (1989), 'The Psychology of Knights and Knaves', *Cognition* 31, pp. 85–116.

29. Rips, Lance J. (1994), *The Psychology of Proof: Deductive Reasoning in Human Thinking*, Cambridge, MA: MIT Press.

30. Rips, Lance J. (1995), 'Deduction and Cognition', in Edward E. Smith and Daniel N. Osherson, eds., *Thinking, 2nd edition*, Cambridge, MA: MIT Press, pp. 297–343.

31. Simon, Herbert A. (1975), 'The Functional Equivalence of Problem Solving Skills', *Cognitive Psychology* 7, pp. 268–288.

32. Stenning, Keith; Cox, Richard; and Oberlander, Jon (1995), 'Contrasting the Cognitive Effects of Graphical and Sentential Logic Teaching', *Language and Cognitive Processes* 10, pp. 333–354.