THE ‘‘UNIVERSAL ACCEPTANCE OF LOGIC’’

By

Christopher Cherniak

Cognitive Science Program

Institute of Cognitive Studies

University of California at Berkeley

February 1983

The production of the Berkeley Cognitive Science Report Series is supported by a grant in cognitive science to the University of California at Berkeley from the Alfred P. Sloan Foundation.

The "Universal Acceptance of Logic"

Christopher Cherniak

Cognitive Science Program

Institute of Human Learning

University of California at Berkeley

February 1983

I

RATIONALITY

The thesis of the universal acceptance of logic is that any rational agent must accept a logic, that is, at least a sound and complete first-order deductive system. I shall argue that the thesis is false under several natural and philosophically important interpretations. The discussion identifies some interesting relations between a theory of non-idealized rationality, computational complexity theory, and recent psychological studies of the formal incorrectness of everyday reasoning. Rejection of the universal acceptance of logic thesis has significant implications, notably for interpretation of Wittgenstein's account of the paradoxes and for Quine's thesis of the universal unrejectability of logic.

First, what is, or would be, the universally accepted logic?[1] At a minimum, the thesis would be that all rational agents accept *some* sound and complete set of laws or axioms and inference rules for first-order logic, as opposed to a claim of the universal acceptance of a particular set of "fundamental" logical laws and rules, or even more strongly, a claim of universal acceptance of all logical truths. The weakest thesis therefore could be false either in that: a) An agent might not accept a complete deductive system; he might have a "cognitive blind spot." Or, b) the agent might only accept an unsound or even inconsistent system; *e.g.*, he might use some rule that did not guarantee preservation of truth in inference, perhaps a "quick and dirty" heuristic. Or more strongly, c) every law or rule the agent used might be unsound or also inconsistent. I shall deal almost entirely with classical logic. This is not to prejudge the issue of the adequacy of non-standard logics; the case of classical logic is a basic one, and the argument should be generalizable to other logics. Whatever one's choice of logic, the prior, and usually unacknowledged, question is whether a sound and complete logic by any standard is in fact the best choice. I also proceed on the simplifying assumption that translation between formal and natural language is generally unproblematic.

[1] It should be noted that at least in more interesting cases, understanding logic involves more than just knowing and using particular logical laws and rules. Even a good logic student has more than a stock of believed laws and rules: He possesses a "metalinguistic" competence to reason about arguments (for example, about their validity); and he has some know-how, in the form of various strategies, for "getting the basic idea" for a proof. There is an important interplay between logic as *post hoc* tribunal in the context of proof, and reasoning (perhaps in terms of mental models) in the context of discovery. However, we concentrate on the simpler matter of just accepting logical laws and rules, a prerequisite in any case for more sophisticated and complex notions of knowledge of logic.

1. Ideal Rationality

Let us examine the concept of a rational agent involved in the universal acceptance of logic thesis. Some rationality constraint on an agent's cognitive system of beliefs and desires is the most fundamental law of psychology, more basic than any low-level empirical generalization. It is generally recognized in the philosophy of psychology, for instance, that while consistency may be the hobgoblin of small minds, consistency is a condition for having any mind at all: no rationality, no agent. The conventional strategy in the cognitive sciences has been first to adopt a very extreme idealization of the rationality required of an agent, and then perhaps, if they are noticed, to explain away departures of real human behavior from the ideal model.

The models of the agent prevalent in decision, game, and economic theory and in philosophy require that the agent is a maximizer of expected utility, that is, that the agent *A* satisfy an ideal general rationality condition, a version of which is: If *A* has a particular belief-desire set, *A* would undertake all and only actions which are apparently appropriate.[2] Here an action is apparently appropriate if, according to *A*'s beliefs, it would tend to satisfy *A*'s desires. An agent who is able to choose his actions so rationally has to have a great deal of logical insight; in particular, he must satisfy an *ideal inference condition*:

> *A* would make all and only the deductively sound inferences from his belief set which are apparently appropriate.

Otherwise, *A* might miss some apparently appropriate actions, for instance. (Nonetheless, in some respects this condition is weaker than the deductive closure conditions customary in epistemic logic.) This logical ability can in turn be characterized as the conjunction of an *ideal appropriateness requirement* on inference choosing ability and an *ideal consequence requirement* on deducing ability, respectively:

i) *A* would select all and only those sound inferences to make from his beliefs that would be apparently appropriate for him to make.

ii) *A* would successfully perform all those inferences.

These last two requirements are interdependent, in that deducing ability is needed to identify a deductive task as currently apparently appropriate, and selecting ability (not necessarily conscious) often is needed to find the deductive sub-tasks that would be helpful in accomplishing a more complex deductive task already undertaken. We shall principally deal with the ideal consequence requirement.

Now, must an ideally rational agent in the sense above accept a logic? For this agent to be able to perform all and only the sound inferences that might turn out to be apparently appropriate he must meet Cartesian standards of perfection: he must in effect both be infallible and have an opinion on everything with respect to logic. Such an agent cannot accept an unsound or incomplete deductive system. If he accepts an unsound system for making some of his inferences, he

---

[2] See my "Minimal Rationality," *Mind* 90 (1981) pp. 161-183. If the requirement that *A only* attempt apparently appropriate actions is dropped, a random guesser will satisfy the ideal conditions, given enough time; see the minimal conditions below.

will not be guaranteed to make only sound inferences, appropriate or otherwise; and if he accepts an incomplete system, he will not be able to make some sound inferences which might turn out to be appropriate. In either case, he does not satisfy the ideal consequence condition. Therefore, an ideal agent must accept a sound and complete logic if he is to perform required reasoning by means of a formal deductive system.

2. Undecidability

If a rationality requirement is the most basic psychological law, the next most basic fact of our psychology is that we are finite objects. The ideal rationality conditions seem incompatible with this fundamental fact of human existence: we are in the finitary predicament of having fixed limits on our cognitive resources, in particular, on our memory capacity and computing time. The standard model seems to assume a human being had God's brain; for such an ideal agent much of the deductive sciences would be trivial. If we suppose the agent is finite, there is still another difficulty for the idea of an ideal agent using only formal deductive procedures. For ideal rationality requires more than just acceptance of a sound and complete logic. The agent has to be able to use that logic very well, so well that any given first-order sentence can be formally proved or disproved in a finite number of steps; otherwise, the ideal agent would not be *guaranteed* always to succeed in making all and only needed sound inferences (*e.g.*, any arbitrary one he thought his survival depended upon). No heuristic procedure for using the deductive system, however good, would suffice unless it was perfect, with no possibility of failure, and hence in fact algorithmic (nor could a recursive enumeration procedure). This finitely represented perfect formal ability, of course, would constitute a decision procedure for first-order logic, which Church's Theorem demonstrates to be impossible.

Hence, the ideal rationality conditions are excluded because they entail either a basic practical impossibility, the use of infinite resources, or else a logical contradiction, like a square circle. Of course, we can still retain the ideal rationality model for some situations as a useful simplification of computational reality, but care is required; doing so now seems a bit as if Hilbert had retained the presupposition of formalism that all number-theoretic truths are formally provable "as a convenient approximation" in the face of Gödel's Incompleteness Theorem.[3]

But perhaps the agent might accomplish deductive tasks by some entirely non-formal means, for instance, by immediate synthetic *a priori* intuition of the deductive relations among the propositions he believed. The agent might do this by direct quasi-perceptual, Gödelian insight into an independent realm of Platonic entities, or by means of his transcendental ego, situated

[3] Undecidability for elementary arithmetic is not known not to arise only for some extremely complex and mathematically uninteresting sentences: See J. Jones, "Three Universal Representations of Recursively Enumerable Sets," *Journal of Symbolic Logic* 43 (1978) pp. 335-351, for an undecidable, unabbreviated sentence (based on the solution of Hilbert's Tenth Problem) of about 100 symbols in length, and his forthcoming paper in the *Journal of Symbolic Logic* for still shorter sentences. (See also G. Chaitin, "Information-Theoretic Limitatons of Formal Systems," *Journal of the Association for Computing Machinery* 21 (1974) pp. 403-424). In addition, a "mathematically simple and interesting" theorem (an extension of the Finite Ramsey Theorem) is not provable in Peano arithmetic; see J. Paris and L. Harrington, "A Mathematical Incompleteness in Peano Arithmetic" in J. Barwise, ed., *Handbook of Mathematical Logic* North-Holland (New York, 1977). The present paper further motivates study of the "density" or distribution of undecidable sentences. See footnote 25 below.

outside of space, time, causality, and so on, as Kant and intuitionists such as Brouwer have identified it. Conformance to the ideal consequence condition through such faculties of intuition may seem little better than doing so by means of an oracle or miraculously perfect luck in guessing; for example, physicalists and those committed to information-processing models of cognition will not be satisfied with even the form of explanations like these. These procedures remain in need of at least the outlines of a scheme of explanation: The alternative procedure has to be one which guarantees inferential success non-algorithmically; furthermore, with intuitive access to such a realm, it is no longer clear that an agent *is* restricted to finite cognitive resources, *e.g.*, of time and space.

3. Computational Complexity

In fact, the deductive ability of the ideal agent is even further removed from computational reality. Even where there is no absolute undecidability, practical undecidability seems to extend further down, to the most basic parts of logic, to the very core of the notion of computation. It is as if, practically speaking, Church's Theorem applied even to the sentential calculus. Of course, a decision procedure exists for tautological soundness of sentential calculus inferences--for example, by use of truth-tables. But while a tautology decision procedure is in principle possible, it now appears to be inherently "computationally intractable," that is, extremely unfeasible as a practical matter, *e.g.*, requiring computations for relatively simple cases that would exceed the capacities of an ideal computer having the resources of the entire known universe. (One of the earliest procedures to be widely recognized as practically unfeasible in this way was exhaustive search of the game tree in chess).

The above tautology result is in the field of computational complexity, an area which has grown rapidly during the last decade and is yielding practical unsolvability results which may be as interesting in some ways for philosophy as the classical absolute unsolvability results of the 1930's.[4] Perhaps philosophers have overlooked the significance of the field so far because of a tendency to conclude that if a problem is decidable in principle, then, at least philosophically, it is trivial.[5] In complexity theory, the feasibility of an algorithm is evaluated in terms of whether its execution time grows as a polynomial function of the size of input instances of the problem: If it does, the algorithm is generally computationally feasible; if it does not and instead increases faster, usually as an exponential function, the algorithm is regarded as computationally intractable. Intractability of a problem turns out to a large extent to be independent of how the problem is represented and the computer model (*e.g.*, random-access or deterministic Turing machine) involved. Just as Turing computability is a formal explication of our intuitive notion of

---

[4] Two key papers are S. Cook, "The Complexity of Theorem-Proving Procedures," in *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing,* Association for Computing Machinery (New York, 1971) pp. 151-158; and R. Karp, "Reducibility Among Combinatorial Problems," in R. Miller and J. Thatcher, eds., *Complexity of Computer Computations*, Plenum Press (New York, 1972) pp. 85-103. A recent review of the field is M. Garey and D. Johnson, *Computers and Intractability*, W. H. Freeman Press (San Francisco, 1979). Two easily accessible articles are H. Lewis and C. Papadimitrious, "The Efficiency of Algorithms," *Scientific American* 238 (1978) pp. 96-109; and L. Stockmeyer and A. Chandra, "Intrinsically Difficult Problems," *Scientific America* 240 (1979) pp. 140-159.

[5] See S. C. Kleene's discussion of the concept of a decision procedure in section 40 of *Mathematical Logic*, John Wiley (New York, 1967) for examples of the relegation of such issues to the applied sciences.

computability, polynomial-time computability can be viewed as a formal specification of a pretheoretic notion of practical computability.

The ideal agent's procedure for determining whether or not a sentence is a tautological consequence of a set of premises yields a test of whether or not a sentence is truth-functionally consistent. In complexity theory, the latter question is known as the "satisfiability problem." Very briefly, the relevant finding here is that the satisfiability problem is a member of the very large and important class of "non-deterministic polynomial time" (NP) problems, ones which are known to be solvable in polynominal time on a non-deterministic Turing machine, which is allowed to make "guesses" and in effect has an unbounded capacity for some parallel computations. But most importantly, the satisfiability problem is "NP-complete": *any* NP problem can be efficiently reduced to the satisfiability problem. In this way, the satisfiability problem is a "universal" NP problem. NP complete problems have not been proven inherently to require exponential time; this is the major unsolved problem, a "Goldbach's Conjecture," of the field. However, NP complete problems are generally regarded as computationally intractable in this way, since only exponential time algorithms for them are known, and because if they were not, so many important problems (such as the Traveling Salesman problem) which have long resisted tractable solution would turn out to be tractable.

Thus, the ideal agent's perfect capacity even just to make all tautologically sound inferences is the case *par excellence* of a problem-solving capacity which is strongly conjectured to require computationally intractable algorithms. Of course, some "quick but dirty" heuristic procedure for tautological inference will not necessarily yield such apparent exponential explosion of computation; presumably that is how actual fallible human beings manage (see sections 9 and 10 below). But again, non-algorithmic procedures would not suffice for the Cartesian perfection of the ideal agent, since *ex hypothesi* they cannot be guaranteed to work in all cases. A surprisingly small fragment of the ideal agent's deducing ability seems by itself to require, for just a finite set of simple cases, resources greater than those available to an ideal computer constructed from the entire universe. There is another layer of impossibility between the idealization and reality.

4. Minimal Rationality

In concluding our discussion of ideal rationality, we can say that the same formalist assumptions needed for the ideal model to imply the universal acceptance of logic thesis make the model unacceptable. While use of an ideal rationality model is an understandable motivation for arriving at the thesis of universal acceptance of logic, some other argument still is needed for that thesis. The idealization strategy seems an over-reaction to the earlier "no rationality, no agent" point. The alternative approach is to begin with a somewhat less idealized, more realistic model, of minimal rationality, where the agent has an ability to choose actions that falls between randomness and perfection.

A minimal general rationality condition would be: If *A* has a particular belief-desire set, *A* would undertake some, but not necessarily all, actions which are apparently appropriate. For an agent to satisfy this condition, he must have some, but not ideal, logical ability, that is, he must satisfy a *minimal inference condition*:

A would make some but not necessarily all the sound inferences from his belief set which are apparently appropriate.

(*A* must also *not* attempt enough of the actions and inferences which are apparently inappropriate; similarly below.) We assume that satisfaction of this condition constitutes having all the logical ability required of an agent. As in the ideal rationality case, the logical ability here can be represented in terms of (*minimal) appropriateness* and *consequence requirements*, respectively:

i') *A* would select some of those sound inferences to make from his beliefs that would be apparently appropriate for him to make.

ii') *A* would successfully perform some of those inferences.

One of the most interesting features of this non-idealized model of rationality is that, as we shall see, it provides a philosophical framework for relating two areas of significant research during the last decade. One is the field of computational complexity just sketched. The other encompasses the many recent psychological experiments that suggest surprisingly ubiquitous use of *prima facie* sub-optimal "heuristic strategies," rather than formally correct procedures, in everyday intuitive reasoning.[6] Although each of these areas has arisen independently of the others, there is a fundamental connection: a) Briefly, complexity theory suggests that human beings (or any computational entities) cannot perform some very simple reasoning tasks in ways that are guaranteed to be correct. b) The psychology of "irrationality" suggests how we can do these tasks, by showing something of how we in fact do them--by means of the "quick but dirty" heuristics. c) The ideal rationality models are at best silent on the normative status of use of these heuristics; the minimal rationality model, to begin with, permits us to acknowledge the basic platitude that human beings are in the finitary predicament, and so *ought* to use some such heuristics--according to this conception, formally incorrect heuristics need not in fact be irrational.

A principal function of this paper, therefore, is to identify further the relations between these three areas. The increasing interest in computational complexity and in psychological heuristics indicates that it is important to establish the status of claims of human (or even inherent computer) alogicality or illogicality. In particular, are the claims somehow *a priori* incoherent and so not a matter open for empirical study, as the rationality idealizations imply? We thereby return to the issue of the universal acceptance of logic. Given the unacceptability of the concept of the ideal agent, the main question now has become: If a supposed cognitive system qualifies as minimally rational, is there any sense in which it must include a logic?

II

[6] For a review of some of their own basic work by two major contributors to the field, see A. Tversky and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science* (1974) pp. 1124-1131. A recent overview is R. Nisbett and L. Ross, *Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice-Hall (Englewood Cliffs, N.J., 1980).

ACCEPTING A LOGIC

5. Strongly Accepting a Logical Law

The most direct argument would be, "An agent must be minimally rational, and minimal rationality in identifying useful actions requires minimal deductive ability. Even just this minimal ability constitutes acceptance of a deductive system; the agent must accept some minimum sound and complete set of basic logical laws and rules in order to perform the required inferences." To determine in what sense, if any, satisfaction of the minimal rationality conditions implies acceptance of a deductive system, we must ask, what is accepting a logical law or rule? The very expression "accepting (or, believing) a logical law" seems something of a philosophical neologism, if not a solecism, but it is a natural concept to introduce and is central to this discussion.

We can distinguish a strong and a weak sense of "accepting a logical law or rule." If *A* accepts a law or rule in the *strong* or realist sense, then *A* would actually use it in making inferences on some of the occasions when this would be apparently appropriate. The key notion here in turn is that of "using a logical law or rule." The related notion of "following a rule" has received much philosophical attention; for our purposes only some relatively straightforward points are needed. Wittgenstein introduces what amounts to this notion early in the *Blue Book*.[7] A rule is "involved" in a process of understanding, obeying, etc. if "the expression [or "symbol"] of the rule forms part of these processes." For a person actually to make use of a rule--one of Wittgenstein's examples is a table of legal chess moves--then would be for that table to be "involved" in this way in the person's act of making a move. Wittgenstein distinguishes this case from a process merely "being in accordance with a rule." A particular chess move that is in accordance with one rule "obviously is in accordance with any number of other rules; and amongst these it is not more in accordance with one than with another."

In fact, for supposed acceptance of a logical law or rule (or indeed any belief) to satisfy the minimal rationality conditions, it must be the above strong acceptance. For, the rationality conditions must be construed counterfactually: Briefly, these conditions require not just that an agent actually *does* act appropriately for a belief *p*, but that he *would* so act, even under circumstances that presently do not obtain. Otherwise, a mere lucky random guesser could satisfy the conditions; such coincidental appropriateness of behavior would be very far from adequate for having a belief system. Now, if the agent would act minimally appropriately for his beliefs under all relevant circumstances, then the ultimate cosmic coincidence is excluded of his not generally doing so in those situations *because* of his beliefs--just as we do not worry about the skeptical possibility that the world was created five minutes ago.

Davidson has noted that a belief must be causally efficacious in the right way in the selection of appropriate actions;[8] to begin with, we can say that the belief must be part of the agent's reason for these actions. (Furthermore, the required causal relation of a belief to appropriate actions undertaken cannot always involve intervening reasoning, or there will be a regress.)[9]

---

[7] Harper and Row (New York, 1958) p. 13.

[8] See, for example, "Psychology as Philosophy," in J. Glover, ed., *The Philosophy of Mind*, Oxford University Press (New York, 1976).

[9] See sections 3 of "Minimal Rationality," *op. cit*. Also relevant is Barry Stroud, "Inference, Belief, and

Now, what is it for an agent to undertake an action in this way because of, among other things, his belief in a logical law or rule--that is, for an agent's belief in a law or rule to part of his reason for an action? It is for the agent to deduce from his other beliefs that the action would best tend to satisfy his desires by appropriate use of the law or rule, which is belief in the above strong sense. Without this constraint of causal efficacy, a point similar to Wittgenstein's applies, that an action is more "appropriate" for one logical law or rule than any other; there is then no distinction with regard to logic between idiot and savant. Granting the open questions regarding what constitutes a reason for an action, and what constitutes really using a logical law at a given stage in a deduction, it is clear in some types of cases that an agent is, *e.g.*, using *modus ponens* and not *modus tollens*. Hence, strong belief or acceptance is required for satisfaction of the minimal rationality conditions.

6. Attributing a Logic

A person's inferences are not always as private as his pain-sensations. In some situations, the inferences of a deducer--whether an agent or not--are indicated by explicit observable performances. An example close to some of Wittgenstein's is a mathematicians's deduction on paper, or as an updated case, the succession of states (under suitable description) in the course of a deduction by a computer running a theorem-providing program. It may then be possible virtually to read off directly how inferences are performed, and in particular what logical laws and rules are being used; in fact, many large computer programs include for debugging purposes a means of directly verifying, by "core dump" on a printout, the sequence of logical operations performed. (For the information-processing conception of psychology, core dump constitutes one sort of model for the ultimate possible knowledge of a human cognitive process.)

However, in many everyday situations it is very difficult for both observer and subject to determine the particular logical laws and rules, if any, the subject is employing. To begin with, as for the above theorem-proving machine, or human natural language processing in general, the user need not use a rule consciously or be able to state, or even understand an explicit statement of, the rule. In addition, even in a very simplified case--say, an agent using a typical textbook natural deduction system consisting of half a dozen rules--the agent might rapidly accumulate a stock of previously proven theorems which he would subsequently employ (not necessarily whenever appropriate) along with the original basic rules. In effect, the agent is then using an enormously redundant deductive system; to find out what "rule" he was using at a given step of a derivation, the observer would have to know a great deal of the past history of the particular agent's use of the original deductive system.

Furthermore, in everyday situations observers in the community usually take for granted a person's (minimal) logical ability. Consequently, when someone asks for the reason why the person undertook an action, it is often not relevant to mention one of the logical beliefs on the basis of which (along with many other beliefs) the agent seems to have decided upon the action. This partly explains why in general, in giving the reason for a person's action, we never even attempt to cite every one of the beliefs involved; we represent the agent's practical inference as an enthymeme. Such a practice conforms to the maxim, "Do not make your conversational contribution

Understanding," *Mind* 88 (1979) pp. 179-196.

more informative than is required," one of the basic maxims H. P. Grice argues are required for efficient or "cooperative" communication.[10] Failure of attributors to mention logical beliefs therefore may reflect the fact that such fundamental beliefs are commonly part of the conversational backgrond, rather than that the agent is in fact not using them. In the deductive sciences such an issue of what logical law a person was using often does move to the foreground. But in everyday practice, instead of attempting to explain or predict a person's deductive behavior by reference to his intervening reasoning processes and the logical laws and rules he uses, we almost always just directly employ a much more easily applied rough profile of his deductive skills--*e.g.*, typically including that inferring p from p & q is easier than inferring $p \rightarrow q$ from $-q \rightarrow -p$. This low-level psychological theory of the difficulty of particular deductive tasks for the deducer in question is in fact entailed by higher-level theories of the agent's reasoning psychology.[11]

7. Weakly Accepting a Logical Law

A challenge to the very notion of using a logical law seems to be implied by Daniel Dennett's account in "The Abilities of Men and Machines"[12] of what it is for a physical object to instantiate or realize a Turing machine. In the course of an attack on anti-mechanists' use of Gödel's Theorem, Dennett argues that it is an error to suppose, "objective and exclusive determinations of the activities and capacities of concrete objects are possible which would determine uniquely which Turing machine specification (if any) is *the* specification for the object." (p. 264) Indeed, in the case of human beings, "By suitable gerrymandering...it ought to be possible to interpret any man as any Turing machine..." (p. 262) If Dennett's claim is extended to physical realizations of abstract algorithms in general, then even our earlier paradigmatically clear case of verifying a computer's logical operations by checking a core dump would be swamped in indeterminacy of interpretation; we could not uniquely describe an object as instantiating a particular computer, much less a computer executing a particular instruction.

But there must be some limits on this indeterminacy of interpretation thesis. Certainly a rock cannot be interestingly or plausibly interpreted as DEC 10 computer system; no one would ever try to use it as a computer. If so extreme a thesis were right, then if the hopes of artificial intelligence turn out to be sound, the rock could be interpreted as a super-computer programmed with human-level intelligence, which would entail a kind of panpsychism. What needs to be shown is that the indeterminacy is not as unrealistic or "hyperbolical" a possibility as the skeptic's challenge that an "evil demon" might have created the world five minutes ago. In everyday life, we would--and do--bet our lives the skeptical possibilities do not obtain. However, this is not to deny that there can be genuine indeterminacy. It can range from small to virtually unlimited; it is significantly present in assertions that a language-user employs a rule tacitly or unconsciously, and can be sometimes so great for a non-linguistic creature like a dog that there

[10] See "Logic and Conversation," in D. Davidson and G. Harman, eds., *The Logic of Grammar*, Dickenson (Encino, Ca., 1975).

[11] See my "Feasible Inferences," *Philosophy of Science* 48 (1981) pp. 248-268.

[12] In his *Brainstorms*, Bradford Books (Montgomery, Vt., 1978). Dennett's indeterminacy of interpretation thesis seems an extension of his claim in "Intentional Systems" (in *Brainstorms*) of a looseness of fit between the "intentional" and "design" levels of description of an object, to an assertion of a corresponding looseness of fit between the "design" and "physical" descriptions involved in the present discussion.

may be virtually no fact of the matter regarding which rule is being used.

The greater the difficulties, practical or conceptual, in determining whether an agent is actually using a particular law or rule, the greater the motivation for employing a *weak* sense of "accepting a logical law or rule": *A* is merely usefully described as using the law or rule in making inferences on some of the apparently appropriate occasions. The weak sense of "accepting a logical law" is derived from the notion of really using a logical law; the strong sense of "accepting a logical law" is the primary one. A description such as, "*A* is behaving approximately *as if* he used *modus tollens*; *in effect A* is using that rule," is characteristic of weak or instrumentalist, as opposed to realist, attribution of acceptance of a rule. Such an attribution may not accomplish much, but it does constitute a step toward the strategy of employing a low-level theory of feasible inferences to account for an agent's deductive behavior. Acceptance of a law is then in the eye of the attributor, rather than an objective matter. Sometimes it helps an observer in explaining an agent's behavior if the observer attributes to the agent acceptance of a deductive system which the agent clearly does not actually employ at all. Even a simple theorem-proving machine that just performs the truth-table method for the sentential calculus could in some situations be usefully described in this way as performing a different procedure of the "natural deduction" type. Claiming the machine at times uses *modus tollens* would be a "rational reconstruction" of the machine's theorem proving processes, a "mere construct" of the attributor, since there would be no event constituting an application of *modus tollens*.

We can now conclude: (i) Assent to a logical law, mere lip-service, is not enough to constitute strong belief in the law. It is neither sufficient nor necessary, although it is one type of evidence for such belief. Also, (ii) acting or reasoning in accordance with, or appropriately for, a logical law is not enough to constitute such belief. (iii) Causal efficacy, the "right" role in the decision-making process, is required; the agent must use the law as a premise in practical reasoning by which he selects aparently appropriate actions.

III

ADEQUACY OF A LOGIC

8. Practical Adequacy

The question of universal acceptance of logic now becomes, must a minimally rational agent accept a logic either strongly or weakly? First, does an agent's satisfaction of the minimal, as opposed to ideal, rationality conditions imply his strong acceptance of a sound and complete deductive system? The question needs further sharpening: One might argue[13] that, while an agent must be able to make some sound inferences--that is, must have some deductive ability--he

[13] As I have in "Feasible Inferences," *op. cit*. See also section 5 of that paper for an argument against the claim that being able to make particular "obvious" inferences, by whatever means, is *constitutive* of understanding the logical constants involved in those inferences; the argument would apply in particular against the assertion that strong acceptance--that is, actual use--of the corresponding obvious valid laws or sound rules is required for an agent to qualify as understanding the logical constants involved.

does not have to be able to make any *particular* inferences, even those that normal human beings find the most obvious. But even if this is true, it does not exclude universal acceptance of logic. It might still turn out that any agent must strongly accept some complete set of valid logical laws and sound rules; it would just be that agents do not have to accept the laws and rules normal humans do--for instance, the ones normal humans find obvious. Our question is therefore, must an agent strongly accept *any* set of valid laws or sound rules that constitute a complete deductive system, much less particular obvious laws or rules? It seems that an agent can have the deductive ability required by the rationality conditions without strongly accepting, *i.e.*, actually using, even one such law or rule.

To begin with, an agent must sometimes select actions without the use of logical laws or rules, or else there will be regresses, as mentioned earlier. For instance, by the appropriateness requirement, an agent must sometimes select apparently appropriate inferences to undertake; otherwise he might entirely squander his limited cognitive resources on completely useless inferences. But even if inferences are therefore reqarded as one kind of action, they cannot themselves always be chosen by means of inferences or the regress begins: "Should I undertake this inference? First I must ask, should I undertake the inference required to settle the previous question?..." Since some of the actions that are inferences cannot be selected by means of reasoning, they cannot be selected by use of a deductive system, sound and complete or otherwise.

In addition, to the extent that one accepts some unproblematic distinction between sentence and image processing,[14] one may hold that there are some cases of human information representation and processing purely in terms of images or sketches. If an agent could select appropriate actions on the basis of his beliefs (in "the right way," and so on) by means of such image processing, this again would not require the use of any deductive system, much less a sound and complete one. However, one objection is that no agent--proprietor of a cognitive system of beliefs and desires--could be entirely an image processor. The argument would be that for a supposed agent appropriately to act on the information represented as an image, he must interpret the image and thereby also be a sentence processor. For, such interpretation requires analysis of features of the image, descriptions (perhaps in the agent's "language of thought" rather than any natural language) which represent those features, sentences to indicate the presence of features, and some deductive system to make needed inferences from those sentences. So, a defender of the universal acceptance of logic could reply that, to qualify as an agent, the image processor must still use logic *sometimes*; indeed, the earlier regress argument is open to a similar point, that an agent's process of selecting actions might still sometimes require the use of logic. I will now restrict consideration to verbally formulated beliefs, and to the question of whether an agent must even sometimes use logic.

For further progress, we need one more distinction. I shall say a deductive system is *metatheoretically adequate* if it is sound (and therefore consistent) and complete. In the first paragraph of "The Justification of Deduction," Michael Dummett asserts, "Failure of soundness yields a situation which must be remedied. Failure of completeness cannot always be remedied;

---

[14] On the debate about the nature and soundness of the distinction, see the papers in Part Two of N. Block, ed., *Readings in Philosophy of Psychology*, vol. 2, Harvard University Press (Cambridge, Ma., 1981). For arguments against the claim that "the brain must represent in sentences" (one of which cites a regress similar to the one discussed in the previous paragraph) see D.C. Dennett, "Brain Writing and Mind Reading," in K. Gunderson, ed., *Minnesota Studies in the Philosophy of Science*, vol. 7, University of Minnesota Press (Minneapolis, 1975), especially p. 410.

a remedy is, however, mandatory whenever it is possible."[15] Dummett describes this absolute requirement of metatheoretical adequacy as "the standard practice of logicians" in constructing and justifying a formal logical theory. However, the metatheoretic adequacy of a deductive system must be distinguished from its *practical adequacy*: here, its adequacy for accomplishing the deductive tasks required of a minimally rational agent. If one assumes any possible agent must be ideally rational, it is easy to overlook the difference between the two types of adequacy. However, with a minimal rationality model, the overlap of the two types of adequacy seems much less salient.

What is the relation between metatheoretic and practical adequacy? I shall point out, first, that practical adequacy does not require metatheoretic adequacy, second, that the former is sometimes preferable to the latter, and third, that the former may not even be compatible with the latter. The very quickest possible, but least reliable, way of performing a deductive task is just to guess the answer. We know a minimally rational agent does not have to be a perfect logician, but the agent could not accomplish his required sound inferences just by a series of lucky guesses (cf. the earlier discussion of counterfactuality). However, there are other ways to improve above chance the odds of selecting conclusions that follow from premises besides using a sound and complete deductive system. The agent might use what is in effect a better than random, but not perfect, gambling strategy for identifying sound inferences. While such a rule of thumb would not always succeed, it might work sufficiently often to reach the break-even point of satisfying the minimal consequence requirement. I shall argue later that this type of strategy may be indispensable if computational intractability is to avoided.

9. Against Metatheoretical Adequacy

The concept of such a strategy suggests to begin with that it is at least possible for a logically competent agent to have one or more "logical blind spots" which are the result of his exclusively using an incomplete deductive system (whenever he does use a deductive system). As an uninteresting example, the agent might use a conventional textbook natural deduction system with a *modus tollens* rule that has a clause that excludes its application to formal sentences with more than 1,000 logical constants. Similarly, it is possible for an agent to perform all of his required sound inferences by means of an unsound, or even inconsistent, system. Frege's axiomatization of set theory in *The Fundamental Laws of Arithmetic* and Quine's axiomatization in *Mathematical Logic* were both inconsistent in ways which did not reveal themselves to many who had used each axiomatization extensively. These two examples suggest that first-order deductive systems can correspondingly be inconsistent in ways which do not yield too many--or any--unsound inferences for the range of deductive tasks required of an agent. Indeed, Kleene notes, "There were defects in all the early formulations of the substitution rule for the predicate calculus";[16] *all* of the early deductive systems for the predicate calculus were unsound in this way. (Of course, if a deductive system is inconsistent it is complete, but this is no longer any kind of a metatheoretic virtue.)

---

[15] In *Truth and Other Enigmas*, Harvard University Press (Cambridge, Ma., 1978) p. 290.

[16] *Op. cit.*, note, p. 107. See A. Church, *Introduction to Mathematical Logic*, vol. 1, Princeton University Press (Princeton, N.J., 1956) pp. 289-290.

Furthermore, an agent who satisfied the minimal rationality conditions could use exclusively a deductive system in which *all* axioms were invalid and *all* inference rules were unsound. A natural deduction system corresponding to a standard textbook one, but composed entirely of unsound rules can easily, if uninterestingly, be constructed. For example, to the original *modus tollens* rule a clause is added: "When one of the premises contains more than 1,000 logical constants, the set of premise-numbers of the line on which the conclusion occurs should be empty; otherwise the premise-numbers are as usually specified." A similar premise-number clause can be added to each of the other rules. Or, the original other rules and the new *modus tollens* rule can just be conjoined as a single rule; such a matter of individuating rules seems arbitrary. The agent would not use the usual shortcut "theorem" rule which permits entering in a derivation a previously proven theorem with an empty set of premise-numbers. The claim here is just that this agent *can* use exclusively this set of unsound basic rules. The agent might happen to be uninterested in using that set to deduce "vacuous" valid sentences; perhaps he has difficulty reasoning so abstractly. As proponents of the naturalness of natural deduction systems often point out, outside of logic courses people rarely seem to use, or at least to cite, logical validities.

The unsound inferences permitted by this system would be performed relatively rarely because they would only arise under a restricted range of conditions: they involve very complex sentences, or are otherwise unintuitive or difficult for the agent to perform. We know an agent cannot perform all inferences--in particular, the more complex ones--anyway, so the unsoundness of this system need not detract *at all* from the agent's rationality. We can conclude that metatheoretically adequate deductive systems are not the *only* way to achieve practical adequacy. One cannot argue that any possible rational agent must accept logic.

Furthermore, it seems likely a stronger point against metatheoretic adequacy holds, that in important cases it is antagonistic to practical adequacy. A metatheoretically inadequate system could be superior to any metatheoretically adequate one for the practical purposes of accomplishing an agent's everyday deductive tasks, just as inconsistent naive set theory is often more convenient than one of the consistent axiomatizations. In such a situation, insisting upon use of a metatheoretically adequate system would itself be unreasonable, like trying to use the more correct but hopelessly unwieldy relativity physics instead of classical mechanics for engineering calculations in designing a bridge. Even outside of practical contexts, empirical theories are often recognized to be idealizations that are only approximations of reality and apply satisfactorily only overlimited ranges of the parameters involved; the kinetic-molecular theory of gases is commonly cited as an idealization that is employed because it is much more manageable than more correct theories, *e.g.*, that do not assume molecules are perfectly dimensionless spheres.[17]

Indeed, as mentioned earlier, a wide range of evidence has recently emerged that indicates human beings do not in fact use formally correct procedures in everyday non-deductive reasoning. And occasionally researchers in this field have pointed out in effect that use of such quick but dirty heuristics in "applied" as opposed to "pure" situations may be a reasonable speed-accuracy tradeoff.[18] There is also a separate long tradition of empirical research suggesting people do not use formally correct procedures in simple deductive reasoning. In addition, my own recent empirical studies suggest subjects use shortcut strategies in deductive reasoning that

[17] The question of adopting convenient but unsound inference rules in fact sometimes is raised in devising natural deduction systems for introductory logic texts; see, for example, B. Mates, *Elementary Logic*, 2nd. ed., Oxford University Press (New York, 1972) p. vii.

[18] See, for example, the last chapter of Nisbett and Ross, *op. cit.*

exploit structuring of concepts in terms of "prototypes", or best examples, of the concepts. This finding is significant because it identifies a connection between heuristics in deductive reasoning and the large range of important recent research on prototype models of mental representation.[19]

10. Practical Paralysis

An even stronger point may hold than just that metatheoretically inadequate systems appear to be preferable sometimes to metatheoretically adequate ones. Some of the recent research on computational complexity raises the possibility that metatheoretical adequacy may in important cases be entirely incompatible with practical adequacy, both for some of the "pure" purposes of the deductive sciences and for the "applied" purposes of maintaining an agent's minimal rationality. Let us begin with an argument of Michael Rabin's[20] for the introduction of a notion of probabilistic proof in mathematics. Of course no decision procedure is possible even in principle for all of elementary number theory, but even such in-principle decidability can sometimes be of very limited value. Consider a result of A. Meyer and L. Stockmeyer's[21] Although the set of theorems of a formal system for the weak monadic second-order theory of successor (S1S), a weakened version of elementary number theory, is decidable in principle, its decidaility is extremely unfeasible in practice. The problem requires not just exponential time, but "super-exponential" time. To prove an arbitrary theorem of only several hundred symbols or less would require a network with so many boolean elements that, even if they were each the size of a proton, the machine would exceed the size of the entire known universe. The "ideal computer" discussed earlier would not suffice. In effect, the moral Rabin drew is that, to avoid the problem of unfeasibly long proofs, mathematicians sometimes should make the ultimate speed-reliability tradeoff: relax the metatheoretic requirement of consistency, even where it is in principle satisfiable, to evade practical paralysis. Rabin recommended and devised methods of probabilistic proof that do not guarantee truth, but for which the probability of error can be determined to be, *e.g.*, one in a billion.[22]

---

[19] An early and subsequently much contested study that suggested subjects use a kind of global impression or "atmosphere" of the logical form of the sentences involved in an inference was R. Woodworth and S. Sells, "An Atmosphere Effect in Formal Syllogistic Reasoning," *Journal of Experimental Psychology* 18 (1935) pp. 45l- 460. See also P. Wason and P. Johnson-Laird, *Psychology of Reasoning*, Harvard University Press (Cambridge, Ma., 1972) ch. 10. On the use of prototypes in deductive inference, see my "Prototypicality and Deductive Reasoning," forthcoming. For a review of earlier research on prototypicality, see E. Rosch, "Human Categorization," in N. Warren, ed., *Studies in Cross-Cultural Psychology*, vol. 1, Academic Press (New York, 1977).

[20] See "Theoretical Impediments to Artificial Intelligence," in J. Rosenfeld, ed., *Information Processing 74*, North Holland Press (Amsterdam, 1974). On Rabin's concept of a probabilistic algorithm, see his "Probabilistic Algorithms," in J. Traub, ed., *Algorithms and Complexity*, Academic Press (New York, 1976); see also C. Pomerance, "Recent Developments in Primality Testing," *Mathematical Intelligencer* 3 (1981) pp. 97-105. Rabin's argument is described in G. Kolata, "Mathematical Proofs: The Genesis of Reasonable Doubt," *Science* 192 (1976) pp. 989-990.

[21] See A. Meyer, "Weak Monadic Second Order Theory of Successor is Not Elementary-Recursive," in A. Dold and B. Eckmann, eds., *Lecture Notes in Mathematics*, no. 453, Springer-Verlag (New York, 1975); and Stockmeyer and Chandra, *op. cit.*

[22] Hilary Putnam has also argued that "quasi-empirical" methods--resembling those employed in evaluating the plausibility, for instance, of highly theoretical statements in physics--have always been important in mathematics outside of the domain of formal proof. See "What is Mathematical Truth?," in *Philosophical Papers*, vol. 1, Cambridge University Press (London, 1975).

Rabin's strategy might be compared with undoing Descartes' bargain; Descartes sought apodeictic certainty, but the cost is generally recognized to have been epistemic paralysis. We can extend the point from methodology of the deductive sciences to our concern about fundamental constraints on human cognition. There are now two possible extremes for dealing with the problem of determining tautological consequence. As pointed out above, just guessing, the quickest but dirtiest procedure, is too dirty for even minimal rationality, since odds of success are chance. The other extreme, of a decision procedure, is the most reliable but also seems too slow; it is perfectly infallible, but it is probably computationally intractable, which might be too slow for even minimal rationality. Therefore, a compromise between the two extremes seems needed to yield sufficient deductive capacity for minimal rationality.

Various tradeoff strategies are in fact prevalent in computer science in dealing with problems that are found to be computationally intractable. Standards are lowered and "heuristic algorithms" and "approximation algorithms" are sought for the problems instead of perfect optimization algorithms.[23] There is more than one type of compromise with perfect algorithmhood that might evade the apparent intractability of decision procedures for even just tautologous consequence. One might, for instance, use a metatheoretically adequate deductive system, but avoid words cases by restricting its application to simpler special cases--sets of premises and a possible conclusion that are small enough so that the exponential explosion of operations does not occur. The agent would never attempt to make an inference from, or to test for consistency, his entire belief set, or even a large portion of it. The cost, and the eventual limitation, is that the agent would then exhibit the most rationality "locally," within certain neighborhoods of his beliefs, and would be particularly weak on inferences and consistency involving beliefs distributed between such subsets. In fact, a fundamental feature of human belief systems is that they are "compartmentalized" in this way; thus, an important rationale for this structuring is as a strategy for evading intractability.[24] Since only some worst cases, as opposed to average cases, of the general problem here are believed to be bad enough to be intractable, a simpler strategy would be: Given any case, employ the metatheoretically adequate system; if no answer results within some fixed time limit, just abandon the attempt and flip a coin to pick an answer.

However, these strategies require that computationally manageable cases involving sufficiently large belief sets are sufficiently frequent to avoid *de facto* computational paralysis. On the other hand, it should be emphasized that the usual complexity measures are worst case, not average case, estimates; the probability distribution of relevant worst cases is not well understood.[25] On the otherhand, some rough estimates do not suggest optimism about methods like the above compartmentalizing and "give up and guess" strategies by themselves being likely to avoid intractability in the management of a human belief system. To begin with, it is of course ominous that for the worst cases, deciding formal sentences of just several hundred symbols in S1S requires more space than there is in the known universe. Presburger arithmetic similarly requires exponential time; furthermore, the exponential explosion of proof length is known to set in early, for sentences of the same order of size as the axiom set. Even decidability of just the monadic

---

[23] See, *e.g.*, ch. 6, Garey and Johnson, *op. cit.*

[24] See my "Rationality and the Structure of Human Memory," forthcoming in *Synthese*.

[25] For an investigation of probability-distribution, as opposed to worst-case, analyses, see R. Karp, "Probabilistic Analysis of Some Combinatorial Search Algorithms," in J. Traub, ed., *op. cit.*. Also relevant for the satisfiability problem is some of B. Dunham and H. Wang, "Towards Feasible Solutions of the Tautology Problem," *Annals of Mathematical Logic* 10 (l976) pp. ll7-l54. The present discussion provides further motivation for research in this area.

predicate calculus (and of some other decidable subclasses of the full predicate calculus) is known to require non-deterministic exponential time.[26]

Let us also consider testing for tautological consequence, only a very small part of the general problem here, by means of the truth-table method (even the more efficient known test procedures such as Wang's algorithm or the resolution method still require as much time in the worst cases). Given the difficulties in individuating beliefs, it is not easy to estimate the number of logically independent atomic propositions in a typical human belief system, but 138 seems much too low. Yet suppose that each line of the truth-table for the conjunction of all these beliefs can be checked in the time a light ray takes to traverse the diameter of a proton, an appropriate cycle time for an ideal computer. At this maximum speed, a consistency test of this very modest belief system would require more time than the estimated twenty billion years from the dawn of the universe to the present.

We should therefore consider an alternative strategy for coping with the computational complexity involved here, the use of a deductive procedure that is not metatheoretically adequate; this type of strategy of course is not incompatible with the two earlier ones. The conclusion that those earlier strategies may not suffice is reinforced in the field of artificial intelligence by recognized practical difficulties for complete resolution-type theorem-proving programs. In general, the algorithmic approach of seeking classical decision procedures, or even just complete proof procedures, for theoremhood needs to be reevaluated in light of the problem of combinatorial explosion of branchings in proofs (some recognition of this difficulty in fact preceded emergence of the field of computational complexity).[27] As in the above cases, the theorem involved here may be extremely simple in comparison with the size even of relevant "compartments" of a human belief system. Especially characteristic of the alternative--selective heuristics--is that they are not complete; as in avoiding exhaustive search in chess, search trees have to be pruned to avoid unmanageable branching, and this prevents proof of some theorems.

And we have seen that a wide range of recent evidence confirms that people actually do not use formally correct methods in their intuitive reasoning. Instead, they seem to use procedures in non-deductive everyday reasoning such as Kahneman and Tversky's "representativeness" and "availability" heuristics, and related procedures in deductive reasoning that exploit the prototype structuring of their concepts. Thus, the most fundamental "empirical" implication of complexity theory, that algorithmic methods of accomplishing even some very simple deductive tasks are likely to be intractable, would provide a unifying framework for explaining why we instead use such heuristics. It constitutes the ultimate "justification of the ways of Man" here. For, a plausible conjecture is that our quick and dirty shortcut strategies are required to avoid intractability. (Of course, this does not exclude an agent from endorsing the use of a metatheoretically adequate deductive system, reasoning *about* it, and perhaps even really using it in some limited contexts.) To the constraint that information-processing models of cognition should be finitary, we could then add that they also should not entail certain classes of computationally complex processes. The possibility needs to be explored further that the only way human reasoning (or any agent's)

---

[26] H. Lewis, "Complexity of Solvable Cases of the Decision Problem for the Predicate Calculus," *Proceedings 19th IEEE Symp. on Foundations of Computer Science (l978) pp. 35-47.*

[27] *For a brief review, see A. Newell and H. Simon, "Computer Science as Empirical Inquiry," Communications of the Association for Computing Machinery* 19 (1976) pp. 113-126. See also Rabin, "Theoretical Impediments to Artificial Intelligence," *op. cit.* An extensive review of practical limitations on theorem provers (and program verifiers) still is needed.

can generally evade practical paralysis is by not using metatheoretically adequate deductive systems. Ideal rationality requirements exclude even entertaining such a possibility; it is in this way that a theory of minimal rationality provides a framework for relating complexity theory and psychological studies of heuristics.

The familiar debate about how to choose between competing logics, *e.g.*, which is more "natural," presupposes as an absolute requirement that we ought not even to consider a logic which is metatheoretically inadequate. But we can now see that the commonplace Dummett began with, that metatheoretic adequacy is mandatory whenever "in principle" possible, needs at least to be carefully restricted. If one is engaged in a metatheoretical investigation of a deductive system, rather than *using* the system, Dummett's assertion may be correct. But if the system is actually to be used significantly, metatheoretic inadequacy may only be repairable on pain of intractability; it would then be extremely inadvisable even to try to adopt a metatheoretically adequate system, since that would preclude successful reasoning. It would be like insisting upon the perfectionism of the Cartesian methodology of universal doubt, with its resulting cognitive paralysis.

11. Universal Weak Acceptance

I have argued that it is possible for an agent not to strongly accept a metatheoretically adequate system, that it may be preferable to accept a metatheoretically inadequate system, and further, that it seems plausible that doing the latter is required to avoid practical paralysis. We can also ask, must an agent weakly accept a metatheoretically adequate system? For example, in "Intentional Systems," Dennett says of adaptively behaving creatures from another planet, "in virtue of their rationality they can be supposed to share our belief in logical truths," and further, of "mice and other animals, in virtue of being intentional systems," "whether or not the animal is said to *believe* the *truths* of logic, it must be supposed to *follow* the *rules* of logic."[28] The context of these remarks confirms that Dennett is dealing with acceptance of logical truths and rules in our weak sense.

But with a distinction between an agent having deductive ability and the agent actually usng a metatheoreticallhy adequate system, even universal weak acceptance of logic does not seem to hold. It is true that an observer can always approximately describe an agent's logical ability by saying that he weakly accepts a metatheoretically adequate deductive system. But it will also always be possible to describe that ability instead by saying the agent accepts a metatheoretically inadequate system. If the latter system is easier for the observer to use, he may even find describing the agent in the former terms gratuitously complex, and so *less* advisable. We have seen that there is no guarantee that the most efficient system--here, for the observer--must be metatheoretically adequate; this is certainly not a matter of a universal *a priori* truth, but of what happens to be convenient for the observer's own current purposes. Therefore, weak as well as strong acceptance of a metatheoretically adequate deductive system is not required for an agent's satisfaction of the rationality conditions. The discussion below concentrates on claims about strong

[28] pp. 9, 11 *op. cit.* My interpretation of these remarks as a claim only of universal acceptance of logic in the weak sense is supported by doubts Dennett expresses in "Brain Writing and Mind Reading" (*op. cit.*; see especially pp. 409-4ll) about the plausibility of an agent's actually using an "extrapolator-deducer mechanism" for inferring new beliefs from a "core library" of beliefs in the form of sentences in brain writing.

acceptance of logic.

IV

IMPLICATIONS

12. Logical Competence

Let us consider consequences of rejection of the thesis of universal acceptance of logic. First, it is sometimes suggested that the logical ability of an agent should be thought of as a "logical competence" on the model of linguistic competence.[29] The fundamental datum of the departures of actual human reasoning from the ideal model is treated as a performance deficit, like those occurring for the linguistic competence of the ideal speaker-hearer as a result of time and memory limitations, etc. There are then two alternative hypotheses to account for actual suboptimal deductive performance: According to the minimal rationality account, human beings might just be using a metatheoretically inadequate deductive system, involving fast but fallacious "heuristics" that are generally practically adequate. According to the competence account, humans have a perfect underlying deductive competence in the form of acceptance of a metatheoretically adequate deductive system; the reasoning errors arise because the logic is applied improperly, as a result of the performance obstructions.

However, is introduction of a concept of competence here really an improvement over talk of knowledge and belief? The notion of a logical competence does not seem to be especially helpful because of types of problems already evident in *Meno* for Socrates' attribution to the servant of knowledge of the Pythagorean Theorem. Briefly, while any hypothesis can be held true come what may, the perfect competence account seems to gain too many epicycles too fast. Eventually, it amounts to saying that an agent accepts a metatheoretically adequate logic; he just usually misapplies it because, in Cohen's words, eliciting conditions are "rarely, if ever, ideal for the exercise of such a competence." (p. 322) Why must the competence be in the form of a metatheoretically, rather than practically, adequate system anyway? A system that is perspicuous from the standpoint of the linguist or logical theorist may not be efficiently processable by the agent. As we have seen, a system that is adequate from the former standpoint may not be processable at all; the computational intractability problems make the competence account particularly implausible. In many ways, the debate here mirrors the complex current debate in psycholinguistics between advocates of elegant "abstract" grammars and advocates of "kluged-up" psychologically real grammars. Indeed, complexity constraints should be germane to both debates. Furthermore, it may prove useful instead to apply the minimal rationality model to some of the difficult problems involved in the competence-performance distinction. In particular, the notion of "accepting a logic" included in the notion of competence here seems to involve equivocation between our strong and weak senses of "acceptance." Does the ideal speaker-hearer, or an actual human being, really use the perfect logic or not?

---

[29] For a recent example, see L. J. Cohen, "Can Human Irrationality be Experimentally Demonstrated?" *Behavioral and Brain Sciences* 4 (1981) pp. 317-370.

13. An "Epistemic Deduction Theorem"

One might argue in another way that an agent must accept some valid laws or sound rules. We know that an agent must be able to make some apparently useful sound inferences. Now, one might argue from the agent's ability to make these inferences to his acceptance of the corresponding logical laws or rules, by appealing to a "quasi-deduction theorem." A version of the deduction theorem of the proof theory of first-order logic is that if ψ is provable from ϕ, then $\phi \rightarrow \psi$ is provable from the empty set, *i.e.*, is a theorem.[30] An epistemic analogue is,

> If a sentence of the form ψ is provable by A at the time t from a sentence of the form ϕ, than A accepts $\phi \rightarrow \psi$ at t.

One would then assert that, where ψ is a consequence of ϕ, if A is able to infer (or, would infer) ψ from ϕ in even any of the "unsound" ways discussed earlier, than A accepts the logical law $\phi \rightarrow \psi$ or the corresponding inference rule. One might also assert that A's inferring ψ from ϕ constitues A's using $\phi \rightarrow \psi$ or the corresponding inference rule, hence in this case A strongly accepts the law or rule. The conclusion is then that an agent must accept logical laws and rules, namely, those corresponding to the inferences he is required to be able to make. The argument implies that an agent accepts more than just a non-redundant deductive system of a few independent rules and axioms; the agent accepts the large set of laws and rules corresponding to the many inferences he is able to make.

However, this "theorem" seems false. As an example, consider a person who frequently undertakes and succeeds in inferring $q \rightarrow -p$ from $p \rightarrow -q$. Suppose that this person in fact always performs this inference and all his other inferences by use of only elementary natural deduction rules such as *modus ponens, modus tollens,* etc. And suppose that at the time the person *would* have similarly performed any other inferences, other things being as equal as possible. That is, it is not the case that he just happens not to think it is useful to perform any inferences in other ways, but would have performed desired inferences in other ways if it has happened to be apparently useful--for instance, if not doing so would have resulted in obviously unfeasibly inconvenient proofs, etc. I concluded earlier that if a person strongly accepts a logical law, he would use it in selecting desirable actions; in particular, he would sometimes use it where it was apparently useful for such desired deductive tasks--for example, when it was practically indispensable for accomplishing them. In the above case, the person does not accept "$(p \rightarrow -q) \rightarrow (q \rightarrow -p)$" or the corresponding inference rule. For, he does not use either of these in performing desired inferences, even when such applications would be apparently useful; and he would not so use them. Instead he uses and would use the elementary deduction rules.

This is clear if we consider a specific instance--the overt and explicit steps written out by a person in performing formal derivations, as opposed to the more problematic steps of his covert reasoning. It is common for a student learning a sentential calculus deductive system to reach a stage of competence at which he behaves as in the above example. That is, when he encounters deductive tasks where using the above contraposition theorem is permissible and appropriate, he instead uses the elementary rules--he cites only those rules as justification at each line of a

[30] Where "ϕ" and "ψ" are variables ranging over sentences of a first-order formal language.

derivation and the lines conform to those rules. Using the contraposition theorem instead of the elementary rules in performing part of a problem might be very valuable (*e.g.*, because of steps and time saved on an examination), and the student may have recognized that using some theorems instead of elementary rules would be very useful. Despite the intricacies of the concept of "using a logical law," in this case it is clear the student is not using the contraposition theorem.

And if the student would consistently behave in this way at this stage of his cognitive history--that is, his failures to use the theorem are not just occasional slips--then he does not then accept the theorem. It is also false that if he performs the inference, he *subsequently* accepts the corresponding law. For, it is common for logic students to have to persuade themselves over and over again of the soundness of an inference, because after each instance of making the inference they do not retain the knowledge that they have successfully performed it. Similarly, when an experienced mathematician has performed a complicated deduction only once as a stage in some proof, he may not have a continuing belief in the associated conditional; for he may immediately forget his earlier result (and rationally do so, since it is unlikely to be useful later), and so have to perform the deduction again if the result again happens to be useful. Therefore, one cannot argue that an agent must accept some valid laws or sound rules by appeal to the claim that being able to infer, or actually inferring, a sentence of the form ψ from one of the form ϕ implies acceptance of $\phi \rightarrow \psi$ or the corresponding rule.

14. The Significance of the Paradoxes

The distinction between metatheoretic and practical adequacy also helps in the interpretation of Wittgenstein's views on the significance of inconsistencies in formal systems such as the semantic and set-theoretic paradoxes. Wittgenstein argues, "One might say, 'Finding a contradiction in a system, like finding a germ in an otherwise healthy body, shows that the whole system or body is diseased.'--Not at all. The contradiction does not even falsify anything."[31] The claim applies to a contradiction whether it is discovered or not: "...as long as [such a contradiction] is hidden, I say it is as good as gold. And when it comes out in the open it can do no harm." (*op. cit*., p. 219) Two theses can be distinguished here: One is that an inconsistent system if *perfectly* good, as good in every way as a consistent system. This view seems to be based on important and difficult anti-realist or constructivist elements in Wittgenstein's philosophy of mathematics: so long as the contradiction has not been "noticed," whether or not the system really is inconsistent is somehow objectively indeterminate. This is the thesis Charles Chihara is concerned with in "Wittgenstein's Analysis of the Paradoxes in His Lectures on the Foundations of Mathematics" when he objects, "the fact that the inconsistency is not noticed will not make it right, any more than not noticing that a person has a cancer will make him well."[32] Among other things, unrecognized contradictions introduce the danger of blindly blundering. According to Wittgenstein, once we do notice a contradiction, we can just isolate it, "not draw any further conclusions from it" (*op. cit*., p. 230). Chihara points out that this maneuver is by no means as trivial as Wittgenstein suggests; there is no guarantee that it can always be effectively performed.

---

[31] *Wittgenstein's Lectures on the Foundations of Mathematics: Cambridge* 1939, C. Diamond, ed., Cornell University Press (Ithaca, 1976), p. 138.

[32] *Philosophical Review* 86 (1977) pp. 365-381; p. 377.

However, we have seen that sometimes it demonstrably can be accomplished. This last point brings us to the less extreme version of Wittgenstein's claim, namely a view argued in this paper, that from the fact that a contradiction can be derived in a deductive system, it does not follow that the system is entirely worthless. The value of a deductive system is not an all or nothing matter; as elsewhere, value here is relative to goals and purposes. Perfect metatheoretic adequacy is required for many investigations in the foundations of mathematics, but we have seen that practical adequacy for some restricted ranges of inferences at a minimum does not always require such metatheoretic adequacy, and indeed metatheoretic adequacy sometimes seems inadvisably costly. Therefore, there is some truth in Wittgenstein's image, "...a contradiction is not a germ which shows a general illness" (*op. cit*., p. 211); the contradiction need not *entirely* vitiate the system. In evaluating the plausibility of the thesis of the universal acceptance of logic, it is important that rejection of this claim be distinguished from the controversial and problematic theses of anti-realism.

15. "Save Logical Truth"

Finally, let us examine Quine's argument for at least the universal unrejectability of logic, based on his analysis of the methodology of translation. Quine's most extensive discussion of this issue is in Chapter 2 of *Word and Object*;[33] he is concerned with the situation in which, on the basis of a person's utterances, an observer tries to determine translations of the sentences the subject accepts. Quine says, "better translation imposes our logic upon [the subject]" (*WO*, p. 58), in that the attributed logical laws must not be contrary to "our logic." Quine states that this claim applies for truth-functional logical laws because, unlike most of language, sentential connectives are not subject to indeterminacy of translation. But Quine's basic thesis is, "fair translation preserves logical laws" (*WO*, p. 59), truth-functional or otherwise; for example, "when someone espouses a logic whose laws are ostensibly contrary to our own, we are ready to speculate that he is just giving some familiar old vocables ('and,' 'or,' 'not,' 'all,' etc.) new meanings." And in a later explanation of the maxim "Save logical truth" in his reply to Barry Stroud in *Words and Objections*, Quine says his *Word and Object* claim was not confined to the tautologies: "I have insisted unconditionally that translation not conflict with any logical truths (*W* & *O* p. 58 ff.)."[34] Correct translation of a subject's utterances should not attribute to him rejection of *any* logical truth. (Presumably this should be generalized to 'Save logical truths and sound inference rules'; that is, correct translation also does not attribute rejection of sound inference rules.)

Now this thesis is weaker than insisting upon universal acceptance of logic. (Although, besides affirming the ambiguous maxims "Fair translation *preserves* logical laws" and "*Save* logical truth," Quine occasionally seems to accept explicitly the latter thesis.)[35] But the argument against the universal acceptance of logic seems to apply even against Quine's thesis of the universal unrejectability of logic. To see this, let us examine the grounds Quine offers for his thesis. Quine refers to the remark

---

[33] MIT (Cambridge, Ma., 1960); hereafter '*WO*'. We concentrate on the views in Quine's earlier writings, through *Philosophy of Logic*, Prentice-Hall (Englewood Cliffs, N.J., 1970).

[34] D. Davidson and J. Hintikka, eds., Reidel (Dordrecht, 1969) p. 317.

[35] See, for example, the last page of *Philosophy of Logic, op. cit*., p. 102.

> that even the most audacious system-builder is bound by the law of contradiction. How is he really bound? If he were to accept contradiction, he would so readjust his logical laws as to insure distinctions of some sort; for the classical laws yield all sentences as consequences of any contradiction. But then we would proceed to reconstrue his heroically novel logic as non-contradictory logic, perhaps even as familiar logic, in perverse notation. (*WO*, p. 59)

Quine's view is that if the putative belief set attributed to a subject includes inconsistencies, for example in the logical laws, the proposed translation of the subject's utterances is incorrect. The above "accepting contradiction" example and much of Quine's other discussion suggests that his translation policy requires the observer to readjust a translation if the translation yields a belief set which is at all inconsistent.

This translation policy presupposes an *ideal consistency condition*:

> If *A* has a particular belief set, then if any inconsistency arose in his belief set, *A* would eliminate it.

(Inconsistencies can be explicit or tacit.) But we do not in fact follow a translation policy of preserving perfect consistency, or anything like it, in attributing a belief set. For example, we know Frege's *Grundgesetze* is inconsistent under customary translations, and we know Frege accepted the axioms of his system. Similarly, Quine's own system in *Mathematical Logic* has turned out to be inconsistent. But neither Quine nor anyone else has suggested that we have misinterpreted Frege or himself, that, *e.g.*, we should abandon the homophonic translation of Quine's axioms. The inconsistency of the axiom sets does not constitute any grounds at all for readjusting these translations. And we can explain this: It would not be rational either for an attributor to adopt so stringent a condition, or for the agent to squander his resources even attempting to satisfy it. And as pointed out earlier, testing merely for truth-functional consistency, the "satisfiability problem," is the example *par excellence* of an NP-complete problem, strongly believed to be computationally intractable. This difficulty for the idealization cannot just be brushed off; even small cases of the satisfiability problem may require resources that exceed the space-time bounds of the known universe.

However, an agent's belief set cannot accumulate inconsistencies indefinitely. It must satisfy a *minimal consistency condition*:

> If *A* has a belief set, then if some but not necessarily all inconsistencies arose in the belief set, *A* would eliminate them.[36]

Assuming an utterer is an agent, this condition does require that correct translation of the sentences the utterer accepts preserve minimal consistency of his beliefs. Clearly, we do not require a translation to eliminate the more difficult tacit inconsistencies; the tolerability of an inconsistency is weighted in terms of its obviousness. Attributors typically employ a theory of feasible inferences which proves this information. A reply at this point is that Quine agrees: While he may not be perfectly consistent himself, given his remark quoted earlier that "translation [must] not conflict with any logical truths," he also asserts in *Philosophy of Logic*, "Save the obvious."

[36] On non-idealized rationality conditions on consistency, see section 5 of "Minimal Rationality," *op. cit*.

(p. 82) Perhaps Quine's principle of charity needs to be applied to itself; Quine must really have had in mind the latter conception of "moderate," rather than perfect, charity. His translation policy then does not conflict with a minimal consistency account.

16. Potential Obviousness

However, Quine seems to take back everything that is granted in "Save the obvious," that is, save only the obvious logical truths, by explicitly and repeatedly equating obvious and unobvious inconsistencies and logical laws. Such failure to distinguish between acceptance of the most obvious and the most difficult logical truths may gain plausibility from the notion of a *theory* employed in the deductive sciences. A theory in this sense is a deductively closed set of sentences; all sentences formally derivable from the set are themselves members of it. No one actually (strongly) accepts every member of such an infinite set of sentences. (No one could even live long enough to read through one of most members of the set, and so on.) But many people of course accept complete deductive systems for first-order logic; they accept a small set of simple axioms and inference rules from which all logical truths could "in principle" be derived. It would not be difficult to confuse such acceptance of a deductive system for first-order logic with acceptance of "the theory of first-order logic": actual acceptance of each of the infinitely many truths derivable by means of those rules and axioms, as opposed to just being committed (or resolving) to accept these truths.

Quine seems to introduce the notion of a virtually logically omniscient agent in this way. He says, "the logical truths are all either obvious [that is, "assented to as a matter of course"] or else potentially obvious, in the sense of being derivable from the obvious by individually obvious steps....We see, then, how it is that 'Save logical truth' is both a convention and a wise one."[37] Quine seems to be self-consciously identifying our actual reasoning process with the idealized derivation process presupposed in proof theory. As a result, he does not distinguish between a theory in the technical sense and complete deductive system for first-order logic; he seems in effect not to be distinguishing the set of assertions of a theory and the set of axioms of the theory. He is then able to conclude "Save logical truth"--*i.e.*, save all assertions of first-order logic---from "Save the obvious"--*i.e.*, save the axioms as he does above.

In *Philosophy of Logic*, Quine does parenthetically qualify this argument: "The canon 'Save the obvious' bans any manual of translation that would represent the foreigners as contradicting our logic (apart perhaps from corrigible confusions in complex sentences)." (p. 83) But the obvious problem for Quine's notion of potential obviousness remains. No finite entity can ever accept, much less actually derive, all of Quine's potentially obvious logical truths, since they constitute the theory of first-order logic--that is, the set of *all* valid sentences, most of which are each too long even to read through. The argument quoted earlier that a person cannot "accept contradiction" because "the classical laws yield all sentences as consequences of any contradiction" suffers from the same problem; we know that no belief set can be deductively closed, and so most of these consequences cannot themselves be in the belief set. Like the rationality idealizatons examined earlier (including the logical competence account), Quine's position entails the

[37] *Words and Objections, op. cit*., p. 318. See also "Carnap and Logical Truth," in *The Ways of Paradox,* Random House (New York, 1966) p. 104; "Epistemology Naturalized," in *Ontological Relativity,* Columbia University Press (New York, 1969) p. 70; and *Philosophy of Logic, op. cit.,* pp. 82-83.

psychological, not just "in principle," triviality of first-order logic, and so of a large portion of the deductive sciences. Yet such theorem-proving does not appear even to be computationally tractable. Human beings can only accept obvious truths and derivation rules, *i.e.*, a deductive system, and derive some of the most "potentially obvious" logical truths. An infinite number of them must remain unknown to us.

17. Moderate Charity

I shall therefore resume examining "Save the obvious" without Quine's own gloss on the principle in terms of potential obviousness. The translation policy is then just: Better translation favors elimination of the more obvious inconsistencies. The next question is, are these inconsistencies to be obvious for the subject or, in translation, for the observer? The principle of charity which Quine says underlies the account of translating logic is that "one's interlocutor's silliness, beyond a certain point, is less likely than bad translation," where this "silliness" is supposed to be evaluated from the observer's perspective.[38] But we cannot assume egocentrically that what is startlingly false for the observer must be startlingly false for the subject. If the observer translates sentences believed by the subject so as to eliminate the inconsistencies the observer finds obviously false, he may not obtain what the subject actually believes; for the subject may be likely to believe those inconsistencies when they aren't obviously false for him. It is not just that this assumption of the universality of obviousness fails to be *a priori* truth. Quite commonplace situations constitute counterexamples: an instructor teaching logic to a new student--what is obvious for the instructor is not for the student--or, Frege's acceptance of his axiom of abstraction before 1903--what is obvious for us today was not for Frege.

Therefore, the translation policy must be: Better translation favors elimination of the inconsistencies which are more obvious for the subject. On the basis of the minimal consistency condition, we still know that some inconsistencies must be obvious for an agent so that he can recognize and eliminate them. To determine what in fact is obvious for the subject, the observer cannot restrict consideration to translation of the subject's utterances; the observer must also take into account the underlying cognitive psychology of the subject. A theory of feasible inferences provides some information on how he reasons, and so which inferences are obvious for him. A theory of the subject's memory structure may also be needed, since a human being can believe even very obviously inconsistent sets of sentences if he happens not to consider them together at the same time. And to find out much about how the subject reasons, organizes his memory, and so on, the observer will need to know something of the particular subject's prior cognitive history.

We can now say that "Fair translation preserves logical laws" must at least be correspondingly revised to: Better translation does not attribute to the subject acceptance of the contradictories of logical truths which are obviously false for the subject. A correct translation policy must allow the possibility that the subject accepts the contradictories of logical truths which are not obviously false for him. But if it is possible the subject finds *no* contradictories of logical truths obviously false, then this translation policy no longer entails even a weak version of the thesis of the universal unrejectability of logic; rather, it is silent on rejectability at least of any particular

---

[38] *Word and Object, op. cit.*, p. 59; see also p. 69.

logical law or rule. (Of course, no one could accept contradictories of all logical laws, any more than he could accept all logical laws, because of the sheer length of most of them.) Now, could a subject find no contradictories of logical truths obviously false? The inconsistencies which, according to the minimal consistency condition, must be obviously false for the agent need not be contradictories of logical laws (the agent may just not be proficient at reasoning abstractly), since not all inconsistent beliefs would be contradictories of logical laws (e.g., "It's raining and it's not raining" is only a particular instance of "*p* & *-p*"). Furthermore, we have seen that accepting and using the contradictories of some logical laws or rules may even be needed for the agent's deductive ability. So far as the rationality conditions are concerned, an agent might find no contradictories of logical laws obviously false and might reject any particular logical laws; in such a case, it would be an error to try to translate him otherwise. Our conclusion regarding even just the universal unrejectability of logic therefore is negative: correct translation might attribute rejection of any particular logical truths.

***

In concluding, let us note the underlying weakness of Quine's account of translation that is relevant here. The defect is that Quine restricts consideration to only translation of the subject's language and determination of the subject's beliefs; thus, for example, Quine's indeterminacy thesis that the subject's language and his beliefs are interconvertible, and so one cannot attribute beliefs without also assigning translations to the subject's sentences, under the constraint of the charity principle. But we have seen that even where Quine grants translation is determinate, an observer cannot ascertain beliefs and translations without a third type of (typically tacit, common sense) theory, of the subject's cognitive psychology. For, the observer cannot apply Quine's charity principle--charitably (re)interpreted--until he has determined what is startlingly false for the subject, and that requires the latter theory.

Given the behaviorist *Zeitgeist* at the time *Word and Object* was published, Quine's tendency to overlook this third factor is itself understandable, if no longer obviously correct to us. Although such a motivation ought not to have force today, Quine's approach is not just his own idiosyncrasy; the conception of philosophy as a "pure" non-empirical discipline seems to reinforce this tendency. As an important example, Davidson[39] and others have espoused a more thorough holism than Quine's by insisting, within a decision-theoretic framework, that desires as well as beliefs and meanings must be considered together. But the holism stops here. The mental--*i.e.*, propositional attitudes and the rationality constraints on them--is supposed to be independent not only of physical reality, but apparently also of psychological reality--of how the agent actually represents and processes information.

However, such a wider theory of interpretation will still require a principle of *moderate* charity, and that in turn cannot be applied at all without taking into account the agent's cognitive psychology, *e.g.*, his memory structure and reasoning processes. It is in this way that a theory of

[39] See especially "Psychology as Philosophy," in *Essays on Actions and Events*, Oxford University Press (New York, 1980). I think the criticism here applies equally to a charity principle which seeks to maximize agreement of subject with observer.

interpretation is constrained to be "psychologically real." To recognize the constraint that theories of propositional attitudes entail only computationally tractable cognitive processes is just a more basic aspect of this turning away form an ultra-autonomist tendency. We have seen that this ultra-autonomism extends still deeper, excluding from consideration even the most fundamental fact of our psychological reality, that we have finite limits on our cognitive resources. The moral regarding the thesis of the universal acceptance of logic is that once the agent's psychological reality begins to be taken into account, that thesis, like a number of other rationality idealizations, seems wrong, and interestingly so. Not only is acceptance of a metatheoretically adequate deductive system not transcendentally indispensable for an agent's rationality, but in important cases it is inadvisable and perhaps even incompatible with that rationality.[40]

[40] This paper was prepared with support from the American Philosophical Society and the National Endowment for the Humanities. Charles Chihara, William Craig, Daniel Dennett, Richard Karp, and Barry Stroud generously helped at various stages of the paper. Some if this material appeared in my Ph.D. dissertation, "Beliefs and Logical Abilities," University of California (Berkeley, 1977). Other material was presented at the Berkeley Logic and Methodology Colloquium, February 1982, and at the Berkeley Philosophy Colloquium, June l982.