Notre Dame Journal of Formal Logic
Volume 34, Number 3, Summer 1993

# Are We Finite?

G. WHITE

**Abstract** This article criticizes the "argument from finiteness"—namely, the argument that, since human epistemic capacities are finite, what we know can be formalized only by using recursively enumerable (and hence first order) theories. Associated with this argument from finiteness is a definition of knowledge as provability in a recursively enumerable theory, and epistemic reasoning as reasoning about provability in such theories. The article proposes an alternative treatment of epistemic reasoning, using the concept of a cognitive state as primitive. It is shown that this concept can be given a well-defined semantics and that this semantics is incompatible with a definition of knowledge as provability in recursively enumerable theories.

***1 Introduction*** This article will be concerned with two things: one is a positive thesis, and the other is a critique. The positive thesis is that we take seriously the idea of a cognitive state, and treat it—in our reasoning about ourselves and each other—as a primitive notion which is not, *prima facie*, susceptible of reduction to other notions. The critique is concerned with a certain argument—I call it the argument from finiteness—which, I believe, hinders our acceptance of the positive thesis. My first task, then, will be to describe the argument from finiteness, and to show why it clashes with the positive thesis, that is, why it tends to prevent us from giving cognitive states the attention they deserve.

***1.1 The critique*** Let us start with the critique, and, in particular, with the argument from finiteness. A particularly clear statement of it runs as follows:

> [many] logical systems . . . have infinitely many theorems—indeed, infinitely many essentially different theorems—but these theorems are finitely generated. There may be infinitely many axioms, but a finite description in the meta-language suffices to specify them. Wffs are finite, and proofs are finite. Such features seem to be inescapable characteristics of logical systems which can actually be used by finite, mortal men. (Andrews [2], p. 239)

*Received April 3, 1992; revised April 27, 1993*

The author is, admittedly, describing a quite particular logical theory, but the characteristics he attributes to it are shared by a very wide class of theories—they are known as recursively enumerable theories, or RE theories for short (cf. Smoryński [20], p. 57); they have many amusing and congenial logical properties, and have been extensively studied. Basically, they are defined by finiteness properties, as detailed above: they should have a set of axioms which can be finitely described, and we should be able to prove theorems from the axioms using, for each theorem, a finite number of steps.

Partly, of course, these theorems have been so extensively studied because they are easy to study, but the more serious reason is the one alluded to above: that they seem to be the only theories which we can conceive of "finite, mortal [women and] men" actually *using*. I shall call this argument the argument from finiteness, and the main conclusion of this paper will be that this argument is specious: that there is more than one way to use a theory, and that most of the philosophically interesting uses do not require theories to be recursively enumerable. But most of all, this argument is misleading because it allows a crucial step to pass unnoticed: one typically first presumes that one's dealings with cognitive states can be treated along the lines of deduction in a logical theory, and only *then* does one apply the argument from finiteness to conclude that such theories must be RE. But, I would argue, it is the crucial first step that has already done most of the philosophically questionable work.

***1.2 The positive suggestion*** Suppose, however, that we *were* to take cognitive states seriously: how would such a theory look? Such things—whatever they might be—could typically figure in two sorts of contexts. On the one hand, many psychological processes have cognitive states as participants: learning, for example, involves adding something to one's cognitive state, whereas forgetting involves losing something from it. On the other hand, when we reason about people's knowledge, we reason about what might, or might not, be in their cognitive state. This is true not only for our reasoning about others' cognitive states, but also for our reasoning about our own—we can speculate, for example, about what we might or might not know tomorrow.

If we are to have a philosophical theory of this sort of thing, then we can either allow cognitive states to occur in it as direct participants, or we can try to reduce them to other things. As I shall argue, most of the work on this sort of area has been reductionist, and this seems to be for two reasons. One is that many of the central problems in the philosophy of mind have to do with the possibility and implications of physicalistic theories of the mind; the elimination of cognitive states, then, can be viewed as part of this program. The other reason is more specifically to do with Wittgensteinian arguments against explanation by means of private mental entities, and the (probably unwarranted) presumption that cognitive states, if they existed, would be examples of such things. This latter reason is more specialized, and I shall not deal with it in this paper. However, the former one is extremely pervasive, and, I would claim, misleading. The debate about physicalism has very much to do with the possibility of *ultimate* reduction, with the final (and probably unattainable) paraphrase of our everyday psychological talk into a purely physicalistic scientific language. This may be all well and good, but our talk of cognitive states is part of our everyday dealings with the world: it belongs together with our talk of tables and chairs. However, even though we know that a complete physicalistic reduction is possible (and, in a sense, trivial) in the case of tables and chairs, they can still be possible objects of philosophical investigation—as, for example, artifacts, as medium sized physical objects, as mereological wholes, as visual objects: and all of these latter sorts of philosophical investigation would, arguably, be considerably impeded if one had to think about physicalism at every turn. However, in the case of the philosophy of mind, things seem to be otherwise: in this case, of course, the question of physicalism just *is* more difficult (and therefore more interesting) than in the case of medium sized physical objects, but, nevertheless, there is no reason why it should always be the main question.

So, then, we shall try to leave cognitive states unreduced (at least proximately), and try to talk about them in their own terms. This will be our rival theory, and it will be in the background when we survey a good deal of work in the philosophy of language, the philosophy of mind, and artificial intelligence. Most of this work tries, implicitly or explicitly, to reduce cognitive states to other entities, and it uses the argument from finiteness while so doing. Thus, a critique of the argument from finiteness has a great deal to do with the project of establishing a rival theory, a theory which takes cognitive states seriously.

## 2 *The main assumption in use*

***2.1 Theories of meaning*** Let us start with the philosophy of language. The basic argument here can be sketched as follows—I follow Platts ([18], pp. 2ff.), although numerous other treatments share the same features. One sets the scene by describing the task of a theory of meaning:

> Suppose a native speaker of some alien tongue emits a string of noises ... What we have to do is make sense of that action; and what that involves is redescribing the action... in such a way as to make [it] intelligible... Such a redescription will come from an overall theory of linguistic behaviour... ([18], p. 2)

Thus, what we are after is a *theory*—by which is meant a theory in the technical sense of formal logic—and this theory will embody both what we know when we understand a foreign language, and what native speakers know when they understand their own language:

> [T]he capacity of finite native speakers to understand a potential infinity of novel utterances... will be comprehensible only if an account of what it is they understand reveals it as deriving from some finite stock of meaning-determining rules and axioms.([18], p. 5)

Proceeding in this way, one soon concludes that the theorems of this theory should include the instances of Tarski's famous Convention T ([18], pp. 6f.).

Of course, appropriate qualifications are usually made. One is not saying, necessarily, that native speakers really have such a theory present in their heads: more that

> we want as close a fit as possible between the manifest competence of the ordinary speaker and that competence that would be manifested by one whose competence was grounded in his explicit knowledge of the theory of meaning, one who had that competence *because of* his knowledge of the theory (as if, so to speak, the latter was the *idealization* of the ordinary speaker.) ([18], p. 15)

The claim, then, is that an agent with knowledge of a RE theory of meaning for the language is a *possible* idealization for the speaker of the language, maybe even a preferred or canonical idealization. And underlying this dogma seems to be the assumption that we can grasp a body of knowledge only if it is (implicitly or explicitly) a RE theory; or even the assumption that we can manifest implicit knowledge in the performance of some practice only if that implicit knowledge can be formalized in a RE theory.

It should be noted here that the possibility of any direct talk of cognitive states has been quietly paraphrased away; and, presumably, one of the perceived advantages of this talk of *theories* is that they seem to be such concrete entities—they are, after all, given by collections of axioms and rules of inference, and all of the latter items can be specified in a reassuringly explicit and down to earth way.

***2.2 The philosophy of mind*** This cluster of assumptions is also at work in a great deal of work on artificial intelligence and the philosophy of mind; we will deal with each in turn. It is important to notice that, in these areas, the argument from finiteness does not occur in isolation, but is generally associated with a group of doctrines, to do with an assumed parallelism between thought and logical deduction and with a program that attempts to eliminate such entities as cognitive states; we will return to these associated doctrines in the last part of this paper, when we develop a philosophical evaluation of the logical material.

***2.2.1 Artificial Intelligence*** Let us start with what has come to be known as "classical AI." This posits mental representations with suitable syntactic structure ("data structures"), and mental mechanisms which operate on these structures according to explicit or implicit derivation rules (cf. Clark [3], p. 286). Such a thing is—generally—described as a "cognitive model," and epistemic terms are quite freely applied to it: Davies, for example, talks of "tacit knowledge" which is embodied in "a causal-explanatory structure [in the brain] which mirrors the derivational structure of the theory" (cited in [3], p. 287). So what is known are theories, with derivational structure, and the causal-explanatory structure of a person's brain must mirror the structure of the theory. And the theories are—because of the finiteness argument—supposed to be RE: not only RE, but one has to be able to search for proofs in them extremely quickly. So in practice one investigates a particular class of theories called Horn theories (cf. van Dalen [23], p. 466); these are the sort of theories you get when you formalize the idea of a set of data structures and derivation rules, and, indeed, they have the advantage of being theories which are very efficient to work with computationally.

This is, of course, not the only view of what AI should be like, but the point over which there seems to be most argument is whether the causal-explanatory structure of the brain should "mirror" the structure of the data. Models are known—so called neural nets—which behave (in favorable cases) exactly like the cognitive models of classical AI, but which do not have the required causal microstructure ([3], pp. 288–293). One can thus investigate neural nets, and still believe that what is known are *theories* (although one does not have to, and a there is a good deal of work devoted to understanding neural nets in their own terms, without considering them as embodying theories ([3], pp. 297–304).

Thus it is that a good deal of work in AI is concerned with the investigation of such theories. A good example is Hayes' project of "naive physics," which attempts

> to construct a formalization of a large part of everyday knowledge of the physical world. Such a formalization could, for example, be a collection of assertions in a first-order logical formalism... I have no particular brief for the usual *syntax* of first-order logic. Personally I find it agreeable: but if someone likes to write it all out in KRL, or semantic networks, or 'fancy' semantic networks of one sort or another, or what have you: well, that's fine... At the level of interpretation, there is little to choose between any of these, and most are strictly weaker than [first order] predicate calculus, which also has the advantage of a clear, explicit model theory, and a well understood proof theory. (Hayes [11], pp. 171,174)

Such a formalization plays two roles: on the one hand, it should be a formalization of our "everyday knowledge," and should thus employ predicates which are recognizably formalizations of concepts that *we* use. On the other hand, it should aid work in AI, so that

> Ideally, one should in principle be able to get a working program from the formalization by assuming a particular inference mechanism and adding *further* information on the meta-level, which 'controls' the inferences this mechanism performs... Looked at in this way, a formalization can be thought of as a 'core' of inferential abilities, whose appropriate deployment at any time for a particular task has to be further specified. ([11], p. 174)

This dual role would account for the restriction to first order predicate calculus (or theories weaker than it): these are all (provided the axiom sets are well behaved) RE theories, and will thus naturally fit into programing work. But this duality is achieved at the price of a certain tension between the two roles; the concepts which most naturally suggest themselves may not fit easily into a RE theory. A good example of this occurs precisely in "naive physics"; a major concern in this area is the behavior of "substances and physical stuffs" ([11], p. 193f.), and of liquids ([11], p. 195). But our naive talk about these things seems to involve talk of *parts*, and the natural way of formalising *this* would seem to be to use higher order logic (or mereology, which is equivalent to monadic second order logic). However, higher order logic is—in general—not RE (cf. van Benthem and Doets [21], pp. 293f.). In this case we have a situation where there is a certain tension between the two aspects: if we try to use the "natural" formalization, we may end up with something that is not RE.

There are other examples of such tensions (see McDermott [15], pp. 221f.), although this one will suffice for the moment. What is worth pointing out is that—despite a good deal of disagreement over details—most authors want a theory to link two things: it should employ concepts that we actually use, but it should also be easily computable—and the requirement of computability plays two roles. Firstly, it means that we can run such models on a computer; but secondly, it means that we can plausibly use the processes as a model of what human beings do when they process data. Thus, such work in AI is trying to achieve, simultaneously, a "content theory" and a "process model" (to use Birnbaum's terminology), where

> A content theory is supposed to be a theory of what people know, how they carve up the world, what their 'ontology' is. A process model explains how they use this knowledge. ([15], p. 222)

It is the combination of these two which gives the argument from finiteness its purchase, since, quite clearly, it is psychological *processes* (and the computations which model them) which must be finite, but, on the other hand, it is *concepts* which we are talking about when we talk about thoughts or beliefs. So that *if* we jump to conclusions and describe thinking as a process of reasoning which gets us from some content to some other content, it seems only natural to describe it as deduction in a RE theory. But even at this stage, we can see that there might be something wrong with these assumptions: there are, namely, tensions, in that sometimes the natural formalization of content may lead us to a seemingly "unnatural" account of the process, an account which cannot be described in RE terms. This suggests that we ought to be attentive to the details of all this; that we should bear in mind the precise logical form of the contents that we are talking about, and about the processes, and that we should check that the two are, in fact, compatible.

***2.2.2 The philosophy of mind proper*** In the philosophy of mind, the focus tends to be different. Here, the emphasis is much more on the metaphysical details of the relation between the physical and the mental; examples tend to be fairly schematic, and most authors—even if they describe the mind as computational—never really specify what sort of computation this is, or what sort of constraints computability imposes on knowledge and reasoning. Thus Peacocke can write, of a mental state corresponding to a disjunctive belief, that

> A device could check that the state which is in fact $S_{Fa \vee Gb}$ is a disjunctive one because it is suitably linked to two belief states; if the system is in this state, and also in one of the states which is the realisation of the negation of one of the states to which it is related in the way characteristic for disjunction (e.g. $S_{\neg Gb}$), the device can cause the system to go into the state $S_{Fa}$. (Peacocke [17], p. 213)

He clearly has in mind some sort of computational model of belief and reasoning, but he hardly goes into any detail about the logical form of the content, or the details of the process: an extremely simple example is taken to suffice. In the passage that I have cited, his emphasis lies elsewhere: on the conflict

> between two... extreme views. One extreme view is an instrumentalist attitude to everyday propositional attitude explanation of someone's actions, the sort given when one uses the scheme of folk psychology. The other extreme view is the claim that there must be a language of thought, and sentence-analogues built up from the vocabulary of this language must in some way be present in the brain of a person with propositional attitudes. ([17], p. 202)

Fodor's position is rather different from Peacocke's, but he still holds that

> it turns out that the philosophical disagreement about whether there is a Language of Thought corresponds quite closely to the disagreement, current among cognitive scientists, about the appropriate architecture for mental models. If propositional attitudes have internal structure, then we need to acknowledge constituency—as well as causal connectivity—as a fundamental relation among mental states. Analogously, arguments that mental states have constituent structure *ipso facto favor Turing/Von Neumann architecture,* which can compute in a language whose formulas have transportable parts, as against associative networks, which by definition cannot.(Fodor [8], p. 139; my emphasis.)

Fodor does talk about content and process, but he does so in an extremely abbreviated way. He takes it that, if mental states have any structured content at all, there is only one candidate for the process model: one ought to consider the brain as a Turing/von Neumann machine, and there is just no plausible alternative ([8], p. 16; cf. Cussins

[5], pp. 378ff.). He does this because he wants to answer the question "What sort of mechanism could have states that are both semantically and causally connected, and such that the causal connections respect the semantic ones?" ([8], p. 14) And his answer is that

> [c]omputers are a solution to the problem of mediating between the causal properties of symbols and their semantic properties. So *if* the mind is a sort of computer, we begin to see how you can have a theory of mental processes that succeeds where—literally—all previous attempts had abjectly failed; a theory which explains how there could be nonarbitrary content relations among causally related thoughts. ([8], p. 19)

And, again, this is a good example of Fodor's interests: he wants to use the concept of computation to answer questions about the relation between semantic and causal properties of mental representations.

Just like the artificial intelligence specialists, the philosophers assume that there is a certain parallel between content and process. The latter do tend to be more metaphysically sophisticated—there is a good deal of enquiry about the question of whether concepts like content actually stand for anything, or whether they are merely theoretical constructs. However, there is a general assumption throughout that, *if* concepts like content stand for things in the brain, then these things in the brain must be capable of being incorporated into a RE model of logical deduction; not only does one have a content/process parallelism, but it is a parallelism which can be described in terms of a particular sort of process (deduction), a process which starts with a particular sort of content (RE sets of axioms).

In this paper, I shall be arguing that such an assumption is vulnerable to the same sort of tensions between content and process that we found in artificial intelligence. However, these tensions are more deeply hidden in the philosophical authors, because of their different emphasis: they are not so much concerned with complex examples, or with the logical minutiae of the problems they are considering. To an extent, this emphasis is perfectly understandable. The philosophy of mind is, after all, concerned with more metaphysical questions, and the philosophical difficulties *seem* to become apparent when one reflects on extremely simple examples: so why make things more complicated? In particular, one can get a good deal of philosophical mileage out of the consideration of one belief at a time, rather than examining the entire system of a person's beliefs; one can similarly go a long way with—let us say—one example of deductive inference and one example of inductive or practical inference; and one can likewise even restrict the logical complexity of the propositions one chooses, by only considering monadic predicates.

It is perfectly understandable, but, I will argue, one also misses something. In Section 3, I shall argue that the tensions which we saw in this section are extremely serious ones: sufficiently serious to make the argument from finiteness invalid. This, in turn, will force a reevaluation of the sort of content/process parallels which we used to set up the argument from finiteness; Section 4 will contain this reevaluation. To conclude this section, it is worth pointing out that the key counterexample to the argument from finiteness is logically rather complex (and, in fact, necessarily so); thus, the philosophers' emphasis on particularly simple examples seems to have led them astray.

***3 The counterexample*** The argument of this section will fall into two stages. First

I shall argue that – if knowledge satisfies a few conditions which seem to be intuitively plausible—we ought to know, *a priori*, a certain complex sentence. The second stage will argue that, if we consider knowledge as provability in a RE theory, we cannot possibly know it.

***3.1 Reflective knowledge of the future*** The main task of this section, then, is to outline a counterexample, which will specify two things: content and process. The argument is not that these are the only candidates for content and process models, but that they are at least coherent and a possible idealization of our common-sense views.

Let us begin with the content. I shall be discussing our knowledge of the future, and for the purposes of this argument I shall think of the future as consisting of an unlimited series of days, stretching from tomorrow onwards. The best idealization will be to consider the days as reaching from tomorrow to infinity: one might think that the range from tomorrow to some arbitrary but very distant future day would be a better idealization, but, as I argue elsewhere (cf. White [26]), if we construe knowledge like *that,* it gives it very counterintuitive properties indeed—properties that make themselves manifest in what is known as the "unexpected examination paradox."

We will consider a sentential logic, augmented with an epistemic operator $K(\cdot)$; if $P$ is any proposition, $K(P)$ will be the proposition 'I know that $P$ is true in future'. For the sake of simplicity, this is the only sort of temporal knowledge that we will consider; we ourselves have much more complex temporal knowledge—for example, we know propositions that will be true only next Tuesday, ones that will be true only on even numbered days, and so on—but we will not be interested in these. Our language—call it $\mathcal{L}$—is thus extremely restricted, but it is all we need for the argument, and this restriction will free us from the necessity for complicated indexing schemes.

We will give this language a semantics by supposing that, on any particular day, any proposition will be true or false; in particular, if any proposition of the form $K(P)$ is true on a particular day, it means that I know *then* that $P$ will be true on all succeeding days. This will give us a way of talking about cognitive states directly; the truth of $K(P)$ on a particular day will mean that the proposition $P$ is a member of the cognitive state corresponding to that day.

We now have to specify the process part of our model. We will, then, suppose that, on any day, we will know something (but not, in general, everything) about which propositions are true, and which false, on certain days in the future. Let us suppose that we can reflect on our knowledge, in the following way: if we know (now, on day zero) that a proposition $P$ is true on days $(i+1)\dots\infty$, which all lie in the future, then we can reflect on this fact, and come to know that $K(P)$ is true on day $i$. Similarly, if we know that $P$ is false on day $i$, we know that $K(P)$ is false on days $1\dots(i-1)$. We also assume that we can reflect in this way now, and on all succeeding days in the future; this is something of an idealization (involving infinitely many steps of reasoning on a particular day), but it is a fairly harmless idealization—it constructs sets of sentences which can be shown to be recursively enumerable, so that *this* part of the idealization is no worse than that of the competing theory. We shall consider that our knowledge of the future is closed[1] under such reflective modification.

Notice several things about this process of reflection. Firstly, it involves deduction on the basis of evidence—but this deduction need not be carried out in the language $\mathcal{L}$; in fact, because $\mathcal{L}$ is so restricted, we are unlikely to be able to do it solely in $\mathcal{L}$. However, for the purposes of modeling mental processes, this will be unlikely to matter provided that this reasoning process can be understood as an activity carried out by a finite entity.

The second thing is that it is necessarily a first person process of reflection. It assumes that I can (now) reflect on what knowledge I have of the future and come to know that $P$ will be true on all days from $i + 1$ onwards; thus, on day $i$, I can come to know that as well, so that, on day $i$, I shall know that $P$ will be true in the future. So, $K(P)$ will then be true. This process of thought, however, uses the identity of the person doing the reflection (now) with the person doing the knowing (on day $i$); it is thus inherently first personal.

***3.1.1 Semantics*** Some sort of semantics is necessary, in order to show that the above model is not contradictory. To this end, we use the concept of a *supervaluation* (cf. van Benthem [22], pp. 45f.): such a thing can be thought of as a partial truth valuation. Formally, we can define a supervaluation to be a set $\Theta = \{\theta_i | i \in S\}$ of (total) truth-valuations $\theta_i$. Then, we define

$$\begin{array}{rll} \Theta(P) = \top & \text{iff} & \theta_i = \top \text{ for all } i, \\ \Theta(P) = \bot & \text{iff} & \theta_i = \bot \text{ for all } i, \\ \Theta(P) & \text{undefined otherwise.} & \end{array}$$

(Note that such a valuation will not, in general, be truth-functional; for example, we can have $\Theta(P \vee Q) = \top$ and both $\Theta(P)$ and $\Theta(Q)$ undefined.) Supervaluations are a representation of a state of partial information; since the reflection process is carried out on the basis of partial information, supervaluations will be an appropriate formalization of it. This will be the content part of our model.

We will also need a process model; that is, we shall be concerned with the process of adding information to a supervaluation. Notice that this makes our cognitive states into dynamic entities; they are not to be identified with fixed supervaluations, but rather the state of each at a particular time will be given by a certain supervaluation, and, as information is added to a cognitive state, the supervaluation will change. This means that, in the language of computer science, they will be objects with state; to quote Abelson and Sussman,

> We ordinarily view the world as populated by independent objects, each of which has a state that changes over time. An object is said to "have state" if its behaviour is influenced by its history. (Abelson and Sussman [1], p. 168)

This may bring our model closer to the real world, but of course it does entail a certain added complexity; Abelson and Sussman write that

> ... no simple model with "nice" mathematical properties can be an adequate framework for dealing with objects and assignment [i.e. entities with state] in programming languages. ([1], p. 175])

This is a rather gloomy assessment, and recent developments have justified rather more optimism.[2] Nevertheless, it does point to the need for conceptual tools with rather more sophistication than the predominantly set-theoretical ones that philosophical logic has used in the past, and which would tend to rule out the possibility of objects with state. (A world in a Kripke model, for example, does not have state, since it simply *is* a set of propositions with certain properties, and, if propositions were added to it, it would by definition be a different world.)

Suppose, then, that we have a supervaluation $\Theta$ (given by a set $S$ of total truth-valuations), and want to add information to it: we have a set $A$ of propositions that we want to make true, and a set $B$ that we want to make false. Consider the set

$$S' = \{v \in S \mid v(P) = \top : \forall P \in A, v(Q) = \bot : \forall Q \in B\}.$$

This gives a supervaluation $\Theta'$ which extends $\Theta$, and is the smallest such which makes all of $A$ true and all of $B$ false. (Note that $S'$ may well turn out to be empty—in this case, $\Theta'$ will make every proposition true and every proposition false. Call this the *trivial* supervaluation. It will, of course, not occur if the sets $A$ and $B$ are consistent with the supervaluation that we started with, which will be the case if the data are partial information belonging to some "real life" situation; in this case, we have a guarantee of consistency.)

On day zero, then, we will have some sort of partial information about which propositions will be true, and which false. Let us represent this by a series of supervaluations, $\langle \Theta_i \rangle_{i \in N}$. On the basis of this, we can carry out the reflection process; that is, if we establish that $\Theta_i(P) = \top$ for $i > n$, we extend $\Theta_i$ to make $P$ true. Similarly, if we establish that $\Theta_n(P) = \bot$, we can extend $\Theta_i$ to make $P$ false for all $i < n$. There is also a converse process of reflection, where we argue from the truth of $K(P)$ at some time to the truth of $P$ at succeeding times; these reflection processes should be thought of as the product of the subject's reflection on the content part of this model of knowledge (they simply draw consequences from the definition that we have given of the epistemic operator $K(\cdot)$).

Further processes of reflection are possible, however. Suppose that we have $\Theta_i(P_1 \wedge \ldots P_k \rightarrow Q)$ for $i > n$. Then the subject could reason as follows: "If I were to gain knowledge, so that I knew that $K(P_1) \ldots K(P_K)$ were true on day $n$, then I would know that the $P_i$ were true after that. But then I could deduce that $Q$ would be true on all days later than $n$; and thus I would also know that $K(Q)$ is true on day $n$." However, a natural representation for the truth of this piece of conditional knowledge is $K(P_1) \wedge \ldots K(P_k) \rightarrow K(Q)$; so it seems natural to add this conditional to the epistemic state of the subject, i.e., to add $K(P_1) \wedge \ldots K(P_k) \rightarrow K(Q)$ to $\Theta_n$. Given the same data, a similar reflection could also add $K(P_1) \wedge \ldots K(P_k) \rightarrow KK(Q)$ to $\Theta_n$. One should notice that these latter two reflection processes come from a reflection on the process part of the model: that is, the subject is here reflecting on the process whereby knowledge is added to his or her epistemic state.

We can tabulate the rules as follows:

| | Data | Effect |
|---|---|---|
| (1) | $\Theta_i(P) = \top, i > n$ | $\Theta_n(K(P)) = \top$ |
| (2) | $\Theta_n(K(P)) = \top$ | $\Theta_i(P) = \top, i > n$ |
| (3) | $\Theta_n(P) = \bot$ | $\Theta_i(K(P)) = \bot, i < n$ |
| (4) | $\theta_i(P_1, \ldots P_k \rightarrow Q) = \top, i > n$ | $\Theta_n(K(P_1) \wedge \ldots (K(P_k)) \rightarrow K(Q)) = \top$ |
| (5) | $\theta_i(P_1, \ldots P_k \rightarrow Q) = \top, i > n$ | $\Theta_n(K(P_1) \wedge \ldots (K(P_k)) \rightarrow KK(Q)) = \top$ |

Here the column labelled 'Data' shows the input to the revision process, and the column labelled 'Effect' shows its output, that is, the truth values that are added to the respective cognitive states.

Notice that these processes are all *monotonic*; that is, they add items to the supervaluations in question but do not take anything away. We can thus iterate the process of reflection until it stabilizes; when this happens, we will have what we shall call a *modally closed* series of supervaluations.

**Definition 3.1** A series of supervaluations $\langle\Theta_i\rangle_{i\in N}$ is *modally closed* iff

$$\begin{array}{rcl} \Theta_i(P) = \top \forall i > n & \Leftrightarrow & \Theta_n(K(P)) = \top, \\ \Theta_j(P) = \bot \text{ for some } j > i & \Rightarrow & \Theta_i(K(P)) = \bot, \\ \Theta_i(P_1 \wedge \dots P_k \rightarrow Q)\forall i > n & \Leftrightarrow & \Theta_n(K(P_1) \wedge \dots K(P_k) \rightarrow K(Q)) \\ \Theta_i(P_1 \wedge \dots P_k \rightarrow Q)\forall i > n & \Rightarrow & \Theta_n(K(P_1) \wedge \dots K(P_k) \rightarrow KK(Q)) \end{array}$$

We can think of these modally closed series of supervaluations as giving a "natural" semantics for our theory: each $\Theta_i$ will represent a cognitive state,[3] and the reflection process describes an operation on cognitive states, which is not *prima facie* given by inference in a theory but by a rule saying directly what happens to the cognitive states in question. The reflection process does, of course, involve some logical deduction, but only in the propositional calculus; there are no such deductions which give any role to the epistemic operators.

Natural though the supervaluation semantics might be, we would still like to know more about it; in particular, we would like to know about the propositions that are true in *all* modally closed series of supervaluations. This is given by

**Proposition 3.2** *A sentence $P$ is true in all modally closed series of supervaluations iff it is a theorem of* **K4DZ**, *the extension of* **K4** *by the axiom schemata*

$$\begin{array}{ll} \textbf{D:} & KP \rightarrow \neg K \neg P \textit{ and} \\ \textbf{Z:} & K(KP \rightarrow P) \rightarrow (\neg K \neg KP \rightarrow P). \end{array}$$

*(I use the terminology of Goldblatt* [10], *p. 51.)*

*Proof:* We should first remark that the set of propositions true throughout every modally closed series is the same as the set of propositions true at the initial supervaluation of every such series; this is clear, because if we take a modally closed series, and remove a (finite) initial segment from it, it remains modally closed.

The only if part is easy to prove by contraposition. Suppose that $P$ is not a theorem of **K4DZ**; then, there is a Kripke model of **K4DZ** which falsifies it. Models of **K4DZ** are trees where all branches have the order type of the natural numbers (imitate the proof on pp. 51 ff. of [10] and leave out references to $\mathcal{L}$); so, we have such a tree—call it $K$—and the root node of $K$ forces $\neg P$. Now, for each node $\nu$ of $K$, let its level $\ell(\nu)$ be the length of the path to it from the root. For each $i$, then, we have a set $\{\nu \mid \ell(\nu) = i\}$ of nodes of $K$; but each node of $K$ determines a total truth-valuation of our language (namely, the valuation which determines which propositions are forced at $\nu$), so for each integer $i$ we have a set of truth-valuations, hence a supervaluation. Call this $\Theta_i$. It is easy to verify that, since we have a model of **K4DZ**, this series is modally closed. However, $P$ is false at the root of $K$, so we must have $\Theta_0(P) = \bot$. Consequently, there is a modally closed series of supervaluations which likewise falsifies $P$.

For the if part, we must show that, if a proposition is a theorem of **K4DZ**, it is true in all modally closed series of supervaluations. Let $\mathcal{T}$ be the set of propositions which are true at the root of every modally closed series of supervaluations. We will first show that $\mathcal{T}$ contains **K4DZ**. To this end, we have to show that $\mathcal{T}$ contains every tautology of the propositional calculus—which is clear, since any supervaluation must validate every such tautology, so, by rule 1, they are all in $\mathcal{T}$. We must next show that it contains every instance of

$$K(P) \wedge K(P \to Q) \to K(Q).$$

However, $P \wedge (P \to Q) \to Q$ is a tautology of the propositional calculus, so must be verified by every $\Theta_i$, so that, by rule 4, $K(P) \wedge K(P \to Q) \to K(Q)$ is verified by every $\Theta_i$. Finally, we must show that $\mathcal{T}$ is closed under the rule of necessitation. Necessitation is easy; suppose that $P \in \mathcal{T}$, then it is true at every valuation of every modally closed series. By rule 5, $K(P)$ must be true at the root node of every modally closed series; so we have $K(P) \in \mathcal{T}$.

It remains to show that the special axioms (i.e. **4**, **D** and **Z**) are in $\mathcal{T}$; this is relatively simple. For any $P$, $P \to P$ is a tautology of the propositional calculus, so that, by rule 5, we have $\Theta_i(K(P) \to KK(P)) = \top$ for any $\Theta_i$ in any modally closed series. So $K(P) \to KK(P) \in \mathcal{T}$, for any $P$; but this is **4**. Since every day has a successor, we have $\neg K \neg P \in \mathcal{T}$, and this is enough to establish **D**. Finally, to establish **Z**, we have to show that

$$K(KP \to P) \to (\neg K \neg KP \to P) \in \mathcal{T};$$

but this follows from the fact that the days have the order type of the natural numbers.

***3.1.2 The sentence*** We are now going to consider a particular example on the basis of this conception of knowledge. As a simplification, we can assume that the future is perfectly known;[4] corresponding counterexamples can be constructed for partially known futures, but this would be a needless complication.

Now let $\perp$ be your favorite absurdity— $0 = 1$, let us say. On any day, this is false in the future of that day, and we know it; thus, by reflection, $K(\perp)$ is false on any day, and hence $\neg K(\perp)$ is true. We know this too. So, on any day, we know that $K \neg K(\perp)$ is true. We also know this. So, by reflection,

$$KK\neg K(\perp)$$

is always going to be true.

***3.2 RE theories*** Suppose that we *can* construe knowledge as provability in a RE theory; so, suppose that, on each day $i$, we have a RE theory $\mathbf{T}_i$ such that $K(P)$ is true on day $i$ iff $P$ is provable in $\mathbf{T}_i$. However, we want more than this: that is, we want to use talk of provablity to *eliminate* talk of knowledge.[5] Now $K(\cdot)$ can be iterated, so we want to be able to iterate talk of provability as well. We know how to do this, of course; by the standard methods of arithmetization, we can enrich our language with a canonical name—$P^*$, say—for each proposition $P$, and, for each of the theories $\mathbf{T}_i$, we can construct a predicate, $Pr_i(\cdot)$, so that $P$ is provable in $\mathbf{T}_i$ iff $P^*$ satisfies $Pr_i(\cdot)$. We can now iterate some of the replacements for the knowledge operators.

But not, at the moment, all of them. The problem is this: a proposition such as $K(P)$ is given a different provability translation on each day, because it is translated using a *different* provability predicate. However, we want to be able to quantify over days; for example, we want to be able to say things like "for all days $i$, proposition $P$ is true at $i$." Such a quantification needs two pieces of syntax; one needs a quantifier on the front of the expression, but one also needs a variable in the expression for the quantifier to match up with. This is fine for nonepistemic expressions, but if we want to quantify expressions involving provability predicates, $Pr_i$, we need these predicates to be constructed in some sort of uniform way, so that the $i$ in them is genuinely a variable and not just a typographic feature. If they are constructed in this way, then we want certain things to hold of the $Pr_i$; for example, if $P$ is a logical truth, we want $\forall i Pr_i(P^*)$ to be provable in arithmetic. (It is clear that it holds for *each individual* $i$, simply because each $Pr_i$ is a provability predicate; however, this can be true without the quantified expression being provable in arithmetic.) In technical terms, we want the $Pr_i$ to be a RE sequence of RE proof predicates. In particular, this will indeed mean that we could use the $i$ in $Pr_i(\cdot)$ as a variable that can be quantified over, and we could—using such quantification—eliminate nested occurences of $K(\cdot)$. I shall show this in action when I talk about the example; the general case is not really difficult, merely rather complicated to describe. Details can be found in [26].

Let us, then, return to the example. The innermost occurence of the knowledge operator is $K(\bot)$; on a given day $i$, this is translated as $Pr_i(\bot)$. When we embed this in the next knowledge operator, we want to say that it is never true on any of the days in the future; so when we evaluate that on day $i$, we will have $Pr_i \forall_{j>i} \neg Pr_j(\bot)$. So, finally, we will have

$$Pr_i(\forall_{j>i} Pr_j(\forall_{k>j} \neg Pr_k(\bot))).$$

But notice that this is stronger than

$$Pr_i(\forall_{j>i} Pr_j(\neg Pr_{j+1}(\bot))); \tag{2}$$

and it follows from the work of Solovay, Smoryński, and Friedman, that if this is true all the $\mathbf{T}_i$ are inconsistent (cf. [20], pp. 179f.). This is not exactly the sort of model of knowledge that one is looking for.

The result of Solovay *et al.* may seem rather counterintuitive, but it is probably best to regard it as an analogue of Gödel's Second Theorem. The latter theorem, remember, says that if a RE theory can prove its own consistency, that theory is inconsistent. (Of course, if a theory is inconsistent, it proves any untruth you want, including $0 = 1$, so there is nothing odd about supposing that it proves that particular untruth which is the assertion of its own consistency.) Solovay's result is a sort of parametrized version of this; it says, roughly, that if you have a chain of theories, and each one proves the consistency of the next one, then every theory in the chain is inconsistent. (As we shall see in the next section, the proof of Solovay's result looks very like the proof of Gödel's Second Theorem; one constructs a predicate, $\rho$, for the whole chain, which has the formal properties of a proof predicate, and one then repeats the reasoning of Gödel's Second Theorem using *that* predicate instead of the usual one.)

***3.3 The technical result*** I shall sketch the proof of the result of Solovay et al.; this section may safely be omitted, but the details of the proof are quite interesting. Suppose, then, that we have a RE sequence of RE proof predicates, $Pr_i$ —corresponding to a sequence of theories $\mathbf{T}_i$ —and suppose that each of the $\mathbf{T}_i$ contains some weak base theory **T** (such as, for example, primitive recursive arithmetic; it need only be strong enough to encode syntax in). Let the proof predicate corresponding to **T** be *Pr*. We want all of the usual properties of proof predicates to hold uniformly in $i$; that is, we want, for any propositions $P$ and $Q$,

$$\mathbf{T} \vdash Pr(P^*) \longrightarrow \forall_i Pr_i(P^*) \tag{3}$$

$$\mathbf{T} \vdash \forall_i \left(Pr_i(P^*) \wedge Pr_i((P \to Q)^*) \longrightarrow Pr_i(Q^*)\right) \tag{4}$$

$$\mathbf{T} \vdash \forall_i \left(Pr_i(P^*) \longrightarrow Pr((Pr_i(P^*))^*)\right) \tag{5}$$

The theorem of Solovay, Smoryński, and Friedman is that, if we have such a sequence of theories, and if **T** proves that each $\mathbf{T}_i$ proves the consistency of $\mathbf{T}_{i+1}$, then all the $\mathbf{T}_i$ are inconsistent; that is, if

$$\mathbf{T} \vdash \forall_i Pr_i((\neg Pr_{i+1}(\bot))^*) \tag{6}$$

then we will have $\mathbf{T} \vdash \forall_i Pr_i(\bot^*)$.

This will establish that, if (2) holds, all the $\mathbf{T}_i$ are inconsistent; we can argue as follows. If **T** is inconsistent, then we are done (since all of the $\mathbf{T}_i$ will be as well). Otherwise, suppose we have (2). Without loss of generality, suppose that $i = 0$, so that we have

$$Pr(\forall_{j>0} Pr_j(\neg Pr_{j+1}(\bot)))$$

By definition of a provability predicate, we thus have (6), so that we now have $\mathbf{T} \vdash \forall_i Pr_i(\bot^*)$; since **T** is *ex hypothesi* consistent, the right hand side must be true, so all of the $\mathbf{T}_i$ must actually *be* inconsistent.

In order to prove that (6) entails what it does, we consider the following predicate:

$$\rho(P^*) = Pr((\forall_i Pr_i(P^*))^*).$$

We can prove (using the above conditions) that it is nicely related to *Pr*, the proof predicate for our "bookkeeping" theory **T**, in the following sense: we can prove that

$$\rho(P^*) \longrightarrow Pr(\rho(P^*)^*) \tag{7}$$

and that

$$Pr((P \leftrightarrow Q)^*) \longrightarrow [\rho(P^*) \leftrightarrow \rho(Q^*)] \tag{8}$$

for any propositions $P$ and $Q$. These two properties—which are extremely weak—imply that $\rho$ must behave like a provability predicate; in particular, $\rho$ must satisfy the following

$$\mathbf{T} \vdash \rho((\rho(P^*) \to P)^*) \to \rho P^* \tag{9}$$

(with *Pr* instead of $\rho$, this is known as "the formalized Löb's theorem.") A proof of this can be found in [20]; it uses little more than known properties of *Pr*, and routine manipulation. Substituting $\bot$ for $P$, we get

$$\mathbf{T} \vdash \rho(\neg(\rho\bot)^*) \to \rho\bot^* \tag{10}$$

Now, suppose that (6) holds; we will now establish that the antecedent of (10) also holds. Thus, so does the consequent; and that was what we wanted to establish. To begin, we need a lemma.

**Lemma 3.3** *Under the above assumptions, we have*

$$\mathbf{T} \vdash \forall i\,(\neg Pr_i\bot \wedge Pr_i\neg Pr_{i+1}\bot \rightarrow \neg Pr_iPr_{i+1}\bot)$$

*Proof:* By (4),

$$\mathbf{T} \vdash \forall i\,(Pr_i\neg Pr_{i+1}\bot \wedge Pr_iPr_{i+1}\bot \rightarrow Pr_i\bot)\,,$$

and we also have

$$\mathbf{T} \vdash \forall i\,(\neg Pr_i\bot \wedge Pr_i\bot \rightarrow \bot)\,,$$

so, by transitivity, we have

$$\mathbf{T} \vdash \forall i\,(\neg Pr_i\bot \wedge Pr_i\neg Pr_{i+1}\bot \wedge Pr_iPr_{i+1}\bot \rightarrow \bot)\,;$$

rearranging a little, we have the result of the lemma.

We can now establish:

**Proposition 3.4** $\mathbf{T} \vdash \forall i\,(Pr_i\neg Pr_{i+1}\bot) \rightarrow \rho\neg\rho\bot$

*Proof:* Suppose, then, that

$$\mathbf{T} \vdash \forall i\,(Pr_i\neg Pr_{i+1}\bot)\,; \qquad (11)$$

by (5), we have

$$\mathbf{T} \vdash \forall i\,(PrPr_i\neg Pr_{i+1}\bot)\,; \qquad (12)$$

since each $Pr_i$ is itself a proof predicate, it satisfies $Pr_iP \rightarrow Pr_iPr_iP$ for any $P$, so we have

$$\mathbf{T} \vdash \forall i\,(PrPr_iPr_i\neg Pr_{i+1}\bot)\,. \qquad (13)$$

Using the lemma on propositions (12) and (13), we have

$$\mathbf{T} \vdash \forall i\,(PrPr_i\neg Pr_iPr_{i+1}\bot)\,; \qquad (14)$$

using the contrapositive of (3), we obtain

$$\mathbf{T} \vdash \forall i\,(Pr_i\neg PrPr_{i+1}\bot)\,; \qquad (15)$$

which (by propositional logic) implies

$$\mathbf{T} \vdash \forall i\,\big(Pr_i\neg Pr\forall j Pr_j\bot\big)\,;$$

finally, we apply (4) to the whole thing and obtain the result.

One could, perhaps, object that this result has nothing really to do with infinite chains of theories, but that all of the action is due to the self-application of the proof predicate for our base theory, **T**, and that this base theory is something imported in order to define our predicate $\rho$, so has nothing to do with the problem at issue. This would be misleading. Observe that the formula (15)—which is where self-application of the said proof predicate first appears—is proved by means of the stronger (14), which has no such self-application for the base theory, but only for the theories that we are studying. Furthermore, this self-application for the proof predicates $Pr_i$ comes from (12) by means of the implication $Pr_iP \rightarrow Pr_iPr_iP$, which is satisfied by any proof predicate at all: it is a simple consequence of demonstrable $\Sigma_1$ completeness (see [20], p. 61). Again, the analogue of Löb's theorem, (9), was proved not by making use of the self-application of some proof predicate, but simply by using the very weak compatibility properties (7) and (8); and it is a feature of the rather peculiar logic of provability predicates that anything which is compatible with a provability predicate must itself behave exactly like a provability predicate. (See [20], p. 173.)

***4 Evaluation*** What does this example tell us? Being a counterexample, its impact might at first seem to be purely negative; however, we can also view it as opening up new territory, as making us aware of distinctions which we had slurred over. We shall, then, try to introduce a few such distinctions—a sort of sketch-map for navigation in new and unfamiliar territory.

***4.1 Concepts of knowledge*** Let us start with a distinction that Cussins makes between two sorts of mental content: this distinction draws on familiar work by Peacocke and others, but it is oriented towards the problems of neurological reduction. One sort—which he calls "$\alpha$ content" ([5], pp. 385–388)—he also describes as "conceptual content"; it is the sort of mental content which can be described by means of concepts—concepts which organize some domain into "objects, properties and situations" ([5], p. 386)—and grasp of these concepts involves grasp of the appropriate truth, or satisfaction, conditions. These contents can be specified by means of a description. And it is $\alpha$ content which is handled by Fodor-style theories:

> ... the level of semantics and the level of syntax are explanatorily independent of each other, in the sense that one does not have to know the semantic theory in order to understand what the theory of the syntax is saying, and vice versa. Syntax must respect semantic constraints, but the operations of proof theory, or procedural consequence, are formal in that they are independent of semantic features. ... We can now see that a semantic theory is a theory of the relation between syntactic items and *conceptual* content. ([5], p. 400)

Similarly, it would be $\alpha$ content that is captured by Davidson's theory of meaning; the descriptions which specify the concepts can be put into Tarski's biconditionals.

There is, however, another sort of content: $\beta$ content. One example of such content is the the indexical 'I'; this "cannot be canonically specified, in the way appropriate to conceptual content, by means of a description" ([5], p. 389). Demonstrative perceptual judgements are similar, as are concepts like 'up' and 'down'. Such conceptual content cannot be accomodated in theories like Fodor's, which rely on being able to separate semantics and syntax, and thereby being able to give a descriptive specification of the mental contents involved:

> ... there is a very large class of cognitive states ... which have a kind of content for which the only canonical conceptual specification is the use of a simple demonstrative or indexical under the conditions of a shared perceptual environment or shared memory-experience ... The problem arises because there is no conceptual structure within the demonstrative or the indexical or the observational content which can be exploited to yield a canonical conceptual specification of the content which would be appropriate for the purposes of a scientific psychology. But this doesn't exclude there being any *nonconceptual* structure within the content." ([5], pp. 302f.)

And psychological investigation of such $\beta$ content will, indeed, involve linking the possession of them to "certain basic, nonconceptual abilities that we possess, such as our ability to move and act in a coordinated way." ([5], p. 399)

Having described Cussins' distinction, we can now ask the question: what sort of conceptual content does our everyday concept of knowledge possess? The usual answer has been that it possesses $\alpha$ content; to spell this out a little, we would think of our concept of knowledge as one of the concepts which organize some domain into "objects, properties and situations."[6] Thus, if we know something, we must know it in virtue of the way in which it belongs to such an organized domain, and "the way it belongs to such a domain" must be capable of being spelled out in terms of whether certain concepts apply to it. Furthermore, these must be concepts with $\alpha$ content, which can be explicitly specified. And thus one is led to think that one's warrant for believing something must be something to do with two sets of $\alpha$ content—the content of the belief, and the content of the concept of knowledge itself—and the way that they interact. So (as Cussins remarks), if one sticks to such content, one is led to "model all cognitive processes as processes of inference." ([5], p. 403) That is, one considers knowledge as, essentially, provability in a theory: this theory is simply a way of making explicit the $\alpha$ content of the concepts involved, and provability is, likewise, a way of making explicit the connection which must obtain between the relevant $\alpha$ contents. And, it seems, the theory must—since we are finite—be a RE theory. As we have seen, this cannot possibly work.

The alternative answer is that our concept of knowledge has $\beta$ content: that is, that it has its role in our mental life by means of being embedded in certain abilities that we have. These abilities will not—as in the case of demonstratives—be abilities to get around in the world, but will be the ability to assess our own cognitive state and that of others: to know when we have a warrant to believe something, to be able to add to (or subtract from) our stock of beliefs, and so on. Let us sketch how one could build up such an account. We could tell an almost Wittgensteinian story, according to which people first learned to make straightforward assertions about the world ("It's raining," "Stanley is asleep," and so on); then they would learn to assess warrants for belief possessed by themselves and others. Maybe this would happen at school: the teacher would ask a question, and then, when the pupil answered, the teacher would say "But do you know?," and so the pupils would learn how to use this new form of words. As I have said, it would be used in the practice of assessing warrants for assertions; and the pupils could, it seems, master this practice without having to assimilate any $\alpha$ content possessed by the concept of knowledge. For example, consider all of the warrants (or counter-warrants) for knowledge which we could in principle recognize. Maybe they would only have a loose sort of family resemblance, so that we could not find a single $\alpha$ content which encapsulated them all. And even if we could sum them all up in a single item of $\alpha$ content, it surely does not follow that this item would encapsulate all the things which we could assert truly and groundedly; this is because our practice of assessing warrants for knowledge is parasitic upon our practice of asserting things (it is to be hoped, truly and groundedly), and not the other way around. There may be many things which we can assert groundedly which we cannot find warrants for, or, if we do have warrants for, cannot extract $\alpha$ content from the warrants. (Consider a dialogue like this: "I think he's lying." "Do you know?" "Yes, I think so." "Why do you think so?" "Something about the way he looked at her. I can't describe it.")

Now because the concept of knowledge belongs with the practice of assessing warrants, it can apply to a very wide range of material; in particular, it can apply to assertions which themselves involve the concept of knowledge. The process of reflection that I described in Section 3.1 forms part of this practice; it is part of the way that we reason about what we might, or could, groundedly assert in the future. As we have seen, it gives very strong constraints on our concept of knowledge; these constraints are strong enough to show that this practice of reflection leads to a concept of knowledge which is perfectly coherent and for which we can specify the appropriate sort of modal logic. And we can show that this concept of knowledge cannot possibly be characterized in terms of $\alpha$ content, if (as seems inescapable) the $\alpha$ content of knowledge is to be spelled out in terms of provability in a RE theory. So the considerations of this section—considerations which might have seemed rather too speculative to have street credibility—can be linked up with a very precise model of the practice of reflection, a model which can be used to prove interesting things about knowledge.

Thus, it is the process of reflection which makes this example interesting. Such processes deserve a good deal of attention; let us indicate why this is so. In this section, I have suggested that the realm of $\beta$ content is the natural home of at least one important philosophical concept (namely, the concept of knowledge), and it is probably the home of many more. It is, very likely, a wild and uncharted domain—it is defined negatively, in terms of what cannot be spelled out in $\alpha$ content but can only be embedded in our practical abilities and our situation in the world. However, the practice of reflection has two important properties: first, it features very prominently and significantly among the things that we do, and, second, it is possible to examine it logically. The practice of reflection, then, could well be a useful clearing in the jungle of $\beta$ content. (Demonstratives, presumably, are another such clearing: and they are the examples that Cussins uses when he develops the distinction.)

Logically, I have said, the concept of reflection can bear investigation. It is connected with what are called "reflection principles": a reflection principle is something like Tarski's *Convention T*—namely, the principle that

$$P \Leftrightarrow P \text{ is true.}$$

Alternatively, reflection principles can be like the principles that we used when we set up the process of reflection—if we ignore the temporal indexing, these ones are of the form

$$\text{I know that } P \Rightarrow P.$$

There are also such principles connected with the concept of provability—they will be of the form

$$P \text{ is provable } \Rightarrow P.$$

Just like RE theories, such things have been the object of logical study for some time; and, just like RE theories, they have amusing and congenial logical properties.[7]

***4.2 Internal and external*** Nevertheless, the argument from finiteness must be good for *something;* we are, after all, finite bits of stuff, and that surely ought to put some sort of constraint upon what we can do. But what sort of constraint? Consider the Wittgensteinian model which we examined earlier, according to which we first learn to make straightforward assertions. Suppose, then, that we are finite in the sense that we can be modeled by finite automata; maybe that our bodily movements can be described as the "output" of such a device. Maybe something even stronger than this is true: for example, conceivably our speech productions (rather than bodily movements) could be considered as the output of a finite automaton. (Partial evidence for this could be the success of Chomsky-style linguistics.) So suppose that we could model the set of sentences which we would (in principle) warrantedly assert by means of some list of sentences which could be produced by some particular Turing machine. And if this list was deductively closed, we could, then, describe it as the list of theorems of a RE theory. Making the list deductively closed seems a permissible idealization, so would not this prove that our knowledge can be described by such a theory?

There is, however, a difficulty. We want to model not only the set of sentences that we know, but also our *concept* of knowledge; we have just shown that there could be a RE theory which is extensionally equivalent to our concept of knowledge, but that does not show that they are the same concept. We need (at least) intensional equivalence for that, and so we need them to behave the same in intensional contexts; in particular, they must behave the same in epistemic contexts. These epistemic contexts are very important, since one of the things that we do with the concept of knowledge is to use it in the process of reflection, and the latter process embeds the concept of knowledge in epistemic contexts of arbitrary complexity. And the argument of this paper has shown that knowledge and provability in a RE theory, however much they may agree extensionally, cannot behave the same in all contexts, whatever the RE theory might be.

We can look at the results of this section in another way. Consider two standpoints, which we may call the internal and the external. The internal standpoint is the one that we ourselves adopt towards our concept of knowledge: we think about the things that we ourselves know, we assess our warrants for them, and so on. It is this concept that is used in the process of reflection, and it is vital that it should be *this* concept; we pass from the fact that we will know $P$ on such and such days in the future to the fact that we ourselves will also know $K(P)$ on another day in the future. It is necessary for the validity of the process of reflection that it should be the same person knowing, and the same concept of knowledge.

There is also another viewpoint, which we can call the external viewpoint. This is the viewpoint of the physicist, or neurologist (*qua* physicist or neurologist, of course[8]); these people consider us folks as finite—if intricate—bits of stuff, and they try to find some elegant and economical description of the things that we do (*qua* bits of stuff). It may be a good idea for such people to model us as finite automata; suppose that it is, and that such research eventually succeeds in describing the set of sentences that we would warrantedly assert to as a list which could be produced by such an automaton. This is, of course, a nontrivial assumption; at the moment we simply do not have such a developed theory, and we do not have any conclusive argument that we ought to be able to attain one. Maybe we are simply too complex to understand ourselves; maybe the goal of a developed neuroscience is in principle unattainable.

But suppose that this assumption pays off, and that we have a developed neuroscience. We will *then* be able to apply the argument from finiteness to the concepts which are at home in that science. However, this will entail nothing at all about our *own* concept of knowledge; all it will have done is to produce a concept which is extensionally equivalent to it, and to which the argument from finiteness applies.

There are, then, two viewpoints, and, *a priori*, they are distinct. But there remains the question of whether these two concepts could *become* the same; of whether, when we have a developed neuroscience, we could choose an appropriate description of ourselves as finite automata, and that, with such a description, the external and internal viewpoints would merge. (People who believe in this story are fond of characterising the realm of concepts like internal knowledge as "folk psychology," and they seem to think that it will be seen to have been related to a developed neuroscience as Galenic medicine is related to modern physiology.) The argument of this paper shows that this cannot possibly happen, so long as reflection remains constitutive of our internal concept of knowledge.

***4.3 Concluding logical postscript*** There are a few things worth noticing about the example. One is that it involves the construction of a *family* of proof predicates. This is no accident; it is very frequent that when we try to parametrize some construction in logic, we find that unexpected difficulties emerge. This seems to be for the following reasons. Quite often, if we naively construct something, we make arbitrary and unwarrantable choices, which we can get away with in an isolated case. However, if we try to do it in the parametrized case, we have to make the choices in a way that depends sensibly on a parameter, which is usually a lot harder, if, indeed, it is possible at all. (Thus it is that proofs of the independence of the axiom of choice involve introducing parameters into the situation, a technique known as forcing.)

So, too, with this example. We could think of it as follows: insofar as the argument from finiteness works, it tells us that, from an external point of view (if such were to exist) we can be considered as finite automata. Suppose we grant this; it is the first arbitrary assumption, since we do not even know if a neuroscience can, in principle, exist. But many philosophers go appreciably further than simply this assumption—they claim that, when we are speaking the truth (or attempting to), what the automaton does is to prove theorems in a RE theory. And there are two questions one can ask here, and which seem to have been given arbitrary answers: firstly, why a *theory*, and secondly, if a theory, which theory, and how is it presented? These are two further arbitrary assumptions.

To see that the first question is genuine, let us think of what the argument from finiteness actually gives us. It would say that when we are speaking the truth, we are producing some sort of list of truths that could be produced by a finite automaton, that is, that is produced by some sort of algorithmic process. This need not be a process of proving theorems in a theory; indeed, there are other procedures known which are extensionally equivalent—i.e., they produce the same list— but which have very different logical properties (they do not satisfy Gödel's second theorem, for example; cf. Detlefsen [6], pp. 344ff.)

Secondly, even if we can assume that what we do is prove theorems in a theory, and even if we know what the theorems of the theory are, there is still the question of what the axioms and rules of inference of the theory should be. This is a crucial point for the philosophy of language and for psychology because—if one is not a hard line instrumentalist—these axioms and rules of inference determine primitive psychological categories which should, in some way, be correlated with goings on in the brain. But the same set of theorems can always be given numerous different axiomatizations, so there is room for a great deal of choice here. With the counterexample that I presented, there are considerable difficulties. The modal logic (**K4DZ**) is certainly a RE theory, and, for each day, the set of things known on that day is a RE theory which contains **K4DZ**. So, day by day, we can axiomatize the knowledge of the future on that day. But the problem is *which* axiomatization; this is vital, because if we change the axiomatization, we change the proof predicate, so that we change what we can prove *about* the proof predicate. And, as we have seen, this matters, because we want to be able to embed proof predicates in each other, so that there are conditions on what we want to be able to prove about the predicates. Conditions which turn out to be impossible to satisfy.

This brings us to the question in the title of this paper. If we are to answer it, we should remember that things are not finite (or indeed infinite) by themselves; things become finite in virtue of certain constructions. Thus, a set becomes finite in virtue of being put into one to one correspondence with a finite set of numbers; a logical theory becomes RE by virtue of being given an axiomatization with certain properties; and so on. From a certain point of view—the external point of view—we may be finite automata. But this is because, from *that* point of view, certain axiomatizations, or descriptions of our activity, become available. However, these axiomatizations cannot be incorporated into our knowledge of ourselves, or (it would seem) into a psychology, however neurologically based that psychology might be: we would be unable to use concepts like knowledge in psychology, if we considered ourselves as finite. What we have is internal knowledge: it is a concept that folks like us have been using for some time, and that seems to serve our purposes fairly well. If we want to venture into philosophical psychology, then conceivably a good way of doing it would be to carry out logical or semantic investigations of concepts like internal knowledge, and, in so doing, to use concepts which are perhaps extensions and corrections of concepts that we naturally use, but, in any case, to take those concepts seriously. One should, I would argue, learn from Kierkegaard's warning (directed against the high theory of *his* time) to remember that "philosophy does not consist in addressing fantastic beings in fantastic language, but [one should remember] that those to whom the philosopher addresses himself are human beings." (Kierkegaard [13], p. 110)

In general, a large part of the business of philosophical psychology must consist of the search for appropriate concepts, concepts which are on the one hand mental and can, on the other hand, be connected with what we know about our physical embodiment and about our own mental life. There are two desiderata here. The first is that it is probably a very good thing if our investigations start from the concepts that we ourselves employ; we will probably not go far wrong thereby, and we are assured of a rich area to start from. As I have attempted to show, it is possible to stay close to these concepts and still to carry out interesting logical investigations. The second desideratum is more negative: it is that these concepts should be naturally, and not arbitrarily, chosen. Every time one chooses an axiom system in an unmotivated way, one is laying up trouble for later. This requirement is quite stringent, especially if one considers situations of any complexity. In particular, parametrization is always a fruitful area for investigation, and not only in logic; there has been a certain amount of work in the philosophy of mind on learning and the acquisition of concepts. ([18], pp. 15f.; [5], p. 408; Clark [4]) But this is not exactly a new thought; it has been a commonplace of the philosophical tradition, since Hegel at least, that if one wants to find out about something, one should not only investigate how it *is,* but also how it changes and develops. And now that we have sophisticated logical tools for studying such situations (see Fourman et al. [9]), it would seem even more to be an important area for investigation. Such an investigation may be technical, but there is no reason why it should not also be faithful to the human situation; as I have attempted to show, these two desiderata can perfectly well coexist.

**Acknowledgment** The author wishes to thank an anonymous referee for helpful criticisms (and thereby a reminder of the author's own finiteness). This research was partially supported by the Swiss National Fund project number 11.31211.91

## NOTES

1. If we want, we can consider a process of iterative modification of our knowledge of the future, in a manner analogous to Kripke's construction (cf. Kripke [14]) of fixed points for a language with truth-predicates. This will start off from a series of partially specified truth values, iterate the process of reflection, and eventually stabilize.
2. For two alternative approaches, see Wadler [24] and [25].
3. Albeit a somewhat complicated one: it is the cognitive state which is given by our knowledge, on day zero, of the cognitive state on day $i$.
4. This will correspond to the model of **K4DZ** which consists of an infinite sequence of worlds with a linear order.
5. This is because talk of RE provability is syntactic, and hence—one might hope—can be linked to some sort of causal story (cf. [8], p. 19). So the elimination of talk of knowledge by talk of provability will be the first step towards the psychological elimination of talk of knowledge.
6. See Section 4.1.
7. A classic paper is Montague [16]. There are good recent treatments of this and related areas by Feferman [7] and Isaacson [12]. Another interesting article is Shapiro [19]; it describes an epistemic logic suitable for the foundations of intuitionism, and uses reflection principles closely allied to those used in this paper.
8. Such people cannot, of course, be considered *qua* their professional roles all of the time; as Kierkegaard points out, the effort to do so would be essentially comic (cf. [13], p. 109).

## REFERENCES

[1] Abelson, H., Sussman, G. J. and J. Sussman, *Structure and Interpretation of Computer Programs*, MIT Press, Cambridge, 1985.

[2] Andrews, P., *An Introduction to Mathematical Logic and Type Theory: to Truth through Proof*, Academic Press, Orlando, 1986.

[3] Clark, A., "Connectionism, competence, and explanation," pp. 281–308 in *The Philosophy of Artificial Intelligence*, edited by M. Boden, Oxford University Press, Oxford, 1990.

[4] Clark, E., "From gesture to word: On the natural history of deixis in language acquisition," pp. 85–120 in *Human Growth and Development*, edited by J. Bruner and A. Garton, Oxford University Press, Oxford, 1976.

[5] Cussins, A., "The connectionist construction of concepts," pp. 368–440 in *The Philosophy of Artificial Intelligence*, edited by M. Boden, Oxford University Press, Oxford, 1990.

[6] Detlefsen, M., "On an alleged refutation of Hilbert's Program using Gödel's First Incompleteness Theorem," *Journal of Philosophical Logic*, vol. 19 (1990), pp. 343–377.

[7] Feferman, S. "Reflecting on incompleteness," *Journal of Symbolic Logic*, vol. 56 (1991), pp. 1–49.

[8] Fodor, J., *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, MIT Press, Cambridge, 1987.

[9] Fourman, M. P., Mulvey, C. and D. Scott, *Applications of Sheaves*, Volume 753 of *Lecture Notes in Mathematics*, Springer–Verlag, Berlin, 1979.

[10] Goldblatt, R., *Logics of Time and Computation*, Vol. 7 of *CSLI Lecture Notes*, CSLI, Stanford, 1987.

[11] Hayes, P., "The naive physics manifesto," pp. 171–205 in *The Philosophy of Artificial Intelligence*, edited by M. Boden, Oxford University Press, Oxford, 1990.

[12] Isaacson, D., "Some considerations on arithmetical truth and the $\omega$-rule," pp. 94–138 in *Proof, Logic and Formalization*, edited by M. Detlefsen, Routledge, London, 1993.

[13] Kierkegaard, S., *Concluding Unscientific Postscript*, translated by D. F. Swenson and W. Lowrie, Princeton University Press, Princeton, 1941.

[14] Kripke, S., "Outline of a Theory of Truth," pp. 53–81 in *Recent Essays on Truth and the Liar Paradox*, edited by R. L. Martin, Oxford University Press, Oxford, 1984.

[15] McDermott, D., "A critique of pure reason," pp. 206–230 in *The Philosophy of Artificial Intelligence*, edited by M. Boden, Oxford University Press, Oxford, 1990.

[16] Montague, R., "Syntactic treatments of modality, with corollaries on reflection principles and finite axiomatizability," *Acta Philosophica Fennica*, vol. 16 (1963), pp. 153–167.

[17] Peacocke, C., *Sense and Content*, Oxford University Press, Oxford, 1983.

[18] Platts, M., "Introduction," pp. 1–18 in *Reference, Truth and Reality: Essays on the Philosophy of Language*, edited by M. Platts, Routledge, London, 1980.

[19] Shapiro, S., "Epistemic and intuitionistic arithmetic," pp. 11–43 in *Intensional Mathematics*, Studies in Logic and the Foundations of Mathematics, North–Holland, Amsterdam, 1985.

[20] Smoryński, C., *Self-Reference and Modal Logic*, Springer–Verlag, New York, 1985.

[21] van Benthem, J. and K. Doets, "Higher order logic," pp. 275–329 in *Handbook of Philosophical Logic*, edited by D. Gabbay and F. Guenther, Reidel, Dordrecht, 1983.

[22] van Benthem, J., *A Manual of Intensional Logic*, CSLI, Stanford, 1985.

[23] van Dalen, D., "Algorithms and decision problems," pp. 409–478 in *Handbook of Philosophical Logic*, edited by D. Gabbay and F. Guenther, Reidel, Dordrecht, 1983.

[24] Wadler, P., "Linear types can change the world!," pp. 561–581 in *Programming Concepts and Methods*, edited by M. Broy and C. Jones, North–Holland, Amsterdam, 1990.

[25] Wadler, P., "Comprehending monads," *Mathematical Structures in Computer Science*, vol. 2 (1992), pp. 461–493.

[26] White, G., "Davidson and an unexpected examination," forthcoming in *Synthese*.

*International Academy for Philosophy*
*Obergass 75, FL–9494 Schaan*
*Liechtenstein*