On the consistency of commonsense reasoning

DONALD PERLIS
University of Maryland, Department of Computer Science, College Park, MD 20742, U.S.A.

Received March 14, 1986
Revision accepted August 25, 1986

Default reasoning is analyzed as consisting (implicitly) of at least three further aspects—oracles, jumps, and fixes—which in turn are related to the notion of a belief. Beliefs are then discussed in terms of their use in a reasoning agent. Next an idea of Israel is embellished to show that certain desiderata regarding these aspects of default reasoning lead to inconsistent belief sets, and that as a consequence the handling of inconsistencies must be taken as central to commonsense reasoning. Finally, these results are applied to standard cases of default reasoning formalisms in the literature (circumscription, default logic, and nonmonotonic logic), where it turns out that even weaker hypotheses lead to failure to achieve commonsense default conclusions.

Key words: beliefs, consistency, introspection, knowledge representation, defaults, circumscription, nonmonotonic logic, commonsense reasoning, ornithology.

Le raisonnement par défaut est analysé comme consistant (implicitement) de trois autres aspects que nous appelons oracles, sauts et fixations, qui sont à leur tour reliés à la notion de croyance. Nous discutons ensuite des croyances en fonction de leur utilisation par un agent qui raisonne. Puis nous relevons une idée de David Israel pour montrer que certains desiderata concernants ces aspects du raisonnement par défaut conduisent à des ensembles inconsistants de croyances et que par conséquent le traitement des inconsistances doit être considéré comme central dans le raisonnement de sens commun. Enfin ces résultats sont appliqués à des cas standards de formalismes de raisonnement par défaut que l'on retrouve dans la littérature (circonscription, logique par défaut, logique non monotone), où il apparaît que même des hypothèses plus faibles ne permettent pas d'obtenir des conclusions de défaut selon le sens commun.

Mots clés: croyances, consistance, introspection, représentation de connaissances, défauts, circonscription, logique non monotone, raisonnement de sens commun.

[Traduit par la revue]

Comput. Intell. **2**, 180–190 (1986)

1. Introduction

Much of the present paper will focus on default reasoning. We will primarily consider a stylized form of default reasoning that appears to be current in the literature, and try to isolate aspects of this stylization that need modification if deeper modelling of commonsense reasoning is to succeed. Specifically, we will show that default reasoning, as has been studied in particular by McCarthy (1980, 1986), McDermott and Doyle (1980), and Reiter (1980) in their respective formalisms (circumscription, nonmonotonic logic, and default logic), will lead to inconsistency under rather natural conditions that we call *Socratic* and *recollective* reasoning. Roughly, a Socratic reasoner is one that believes its default conclusions in general to be error-prone, and a recollective reasoner is one that can recall at least certain kinds of its previous default conclusions. We will show that the standard approaches, based on what we term *jumps* (as in *jumping to a conclusion*), are inconsistent with these desiderata.[1]

This is not to say that research into these formalisms has been misguided, or that their authors have assumed that they were adequate for all contexts. On the contrary, these studies have been essential first steps into an area of high complexity demanding a "spiralling" approach of more and more realistic settings. Here then we hope to show one further stage of development that is called for. In fact, elsewhere (Drapkin *et al.* 1986) we have argued that inconsistency is a somewhat normal state of affairs in commonsense reasoning, and that mechanisms are needed for reasoning effectively in the presence of inconsistency.

Note that there are at least two general frameworks in which such formal studies can proceed: We can seek a specification of formal relations that might hold between axioms and inferences in a supposed default reasoner (the "spec" approach), or we can seek to identify specific actions that constitute the process of drawing default conclusions (the "process" approach). That these are related is no surprise. In effect, the first is a more abstract study of the second, aimed at providing a characterization of what sort of things we are talking about in our study of defaults. However, there is a hidden further difference, namely that in pursuing the former, one is naturally led to consider idealized situations in which features irrelevant to the particular phenomenon at hand are deliberately left out of consideration. Such an approach has been customary in research in commonsense reasoning, most conspicuously in the assumption of logical omniscience: that an agent knows (and even instantly) all logical consequences of its beliefs, generally regarded as part of the notion of epistemological adequacy. That is, although no one believes that agents *actually* can reason this way, it has seemed to be a convenient test-bed for ideas about what reasoning is *like*, apart from the "noise" of the real world.

While this has come under criticism lately, and while authors of default formalisms acknowledge the importance *to their very topic* of the process nature of defaults,[2] still the latter has remained conspicuously absent from the continuing development of such formalisms. Here we argue that the very essence of

[1] "Jumping to conclusion can lead to unpleasant landing." Chinese fortune cookie, 1986.

[2] At least regarding their context of an overall process of reasoning going on over time within a changing environment of inputs. For example, McDermott and Doyle (1980, p. 41) speak of "... modelling the beliefs of active processes which, acting in the presence of incomplete information, must make and subsequently revise assumptions in light of new observations." Reiter (1980, p. 86) mentions "... the need for some kind of mechanism for *revising beliefs* [his emphasis] in the presence of new information. What this amounts to is the need to record, with each derived belief, the default assumptions made in deriving that belief. Should subsequent observations invalidate the default assumptions supporting some belief, then it must be removed from the data base."

default reasoning, and of commonsense reasoning in general, derives from its being embedded in the real world, and in agents evolved to deal with such by means of an appropriately introspective view of their own fallibility and corrigibility over time. This in turn will be seen to pose problems for logical omniscience. We refer to the "spec" view of an ideal thinker, which we are critiquing, as that of an "omnithinker" (or OT for short).

To facilitate this discussion, we first present an extended illustration of default reasoning along lines found in the literature. Reasoning by default involves reaching a conclusion C on the basis of lack of information that might rule out C.[3] For example, given that Tweety is a bird and no more, one might, if prompted, conclude (at least tentatively) that Tweety can fly.[4] Here C is the statement that Tweety can fly. Such a conclusion may be appropriate when "typical" elements of a given category (in this case, birds) have the property under consideration (ability to fly).

In the bird example, additional information, such as that Tweety is an ostrich or that Tweety has a broken wing, should block the conclusion that Tweety can fly. Just how it is to be determined that the conclusion is or is not to be blocked is still a matter of debate. Nonetheless, we can usefully discuss the phenomenon of default reasoning in the form of a sequence of steps in a suitable formal deduction as follows: We simply list the "beliefs" that a reasoning agent may consider in drawing the default conclusion. This will be done first in a very sparse form, and then in a more amplified form. The following list is intended to be in temporal order as our reasoner "thinks" first one thought (belief) and then another. Step (2) is the usual default conclusion, given belief (1), and steps (3) and (4) illustrate the apparent nonmonotonicity in which additional data (3) can seemingly block or contradict an earlier conclusion.

(1) Bird(Tweety) [this simply comes to mind, or is told to the agent]
(2) Flies(Tweety) [this "default" comes to mind, perhaps prompted by a question]
(3) Ostrich(Tweety) [observed, remembered, or told to the agent]
(4) $\neg$ Flies(Tweety) [this comes to mind]

Now, the above sketch of reasoning steps ignores several crucial points. In particular, the direct passage from (1) to (2) obscures several possible underlying events,[5] namely, (a) a recognition that it is not known that Tweety cannot fly (i.e., Tweety is not known to be atypical regarding flying), (b) an axiom to the effect that flying is indeed typical for birds (so that if a bird can be consistently assumed typical, it is likely that it can fly, and therefore a reasonable tentative assumption), and (c) the willingness to go ahead and use the tentative assumption as if it were true.[6]

Similarly, the passage from (3) to (4) obscures the necessary information that ostriches cannot fly. Finally, (4) leaves unsaid the implicit conclusions that (a) Tweety must be atypical and (b) the assertion that Tweety can fly is to be suppressed. The following sequence illustrates this more explicit description.[7] Certain of the new steps have been labelled with the terms *oracle*, *jump*, and *fix*. For the moment, we limit ourselves to the following brief remarks (later we will elaborate on them). Oracles are what make default reasoning computationally difficult and also what make it nonmonotonic; jumps are what make it shaky (unsound); and fixes keep it stable. Only the work of McCarthy (1980, 1986) has seriously addressed the oracle problem; McDermott and Doyle (1980) and to some extent Reiter (1980) have characterized jumps; and Doyle (1979) has the only work on fixes.

(1) Bird(Tweety) [axiom]
(1*a*) Unknown($\neg$Flies(Tweety)) [oracle]
(1*b*) Unknown($\neg$Flies(x)) & Bird(x) .$\rightarrow$ Tentative(Flies(x)) [default]
(1*c*) Tentative(Flies(x)) $\rightarrow$ Flies(x) [jump]
(2) Flies(Tweety) [consequence of 1, 1*a*, 1*b*, 1*c*]
(3) Ostrich(Tweety) [new axiom]
(3*a*) Ostrich(x) $\rightarrow$ $\neg$Flies(x) [axiom]
(4) $\neg$Flies(Tweety) [consequence of 3, 3*a*]
(4*a*) Flies(x) & $\neg$Flies(x) .$\rightarrow$. Suppress(Flies(x)) & Atypical(x) [fix]
(4*b*) Atypical(Tweety) [consequence of 2, 4, 4*a*]
(4*c*) Suppress(Flies(Tweety)) [consequence of 2, 4, 4*a*]

To restate quickly (and a bit oversimply) our aims in this paper, we will argue that any mechanism for default reasoning that utilizes jumps will be inconsistent if it also is Socratic (believes it can make mistakes) and recollective (recalls its past conclusions). Moreover, the above expanded scenario with fixes strongly suggests the need for precisely these kinds of additional features (Socratic and recollective).

It is now time to turn to an extended look at the nature of beliefs, since the cursory treatment of defaults above should make clear that beliefs are the stuff that defaults are made of, and that if we do not know what beliefs are, at least in rough form, then we will remain in the dark about defaults as well. More specifically, addressing the issue of *consistency* of an agent's set of beliefs makes it essential to decide, at least informally, what counts as a belief.

2. A preliminary analysis of beliefs

Much AI literature purports to be about the beliefs of reasoning agents, e.g., Moore (1977), Perlis (1981), and Konolige (1984). Yet little in this literature has been said as to what actually makes something a belief.[8] While it is acknowledged that the ontological character of beliefs is unclear, and that at least two approaches are worth considering (the syntactic and the propositional), not much attention has been given to the issue of what distinguishes a statement or proposition that is a belief from one that is not. Agents are endowed with a fairly arbitrary set *Bel*, subject perhaps to the requirements of being

[3]Doyle (1983, 1985) has presented interesting views on this phenomenon, relating it to group decision making.

[4]Tentativity is the obvious key, which all default formalisms are designed to capture, as opposed to other more robust kinds of inferences. And it is this that our analysis will focus on most.

[5]Here emphasis is placed on "possible." It is not claimed that these events must occur; but we will argue that for certain commonsense situations they are appropriate.

[6]Note that Nutter (1982, 1983*a*) in effect cautions against careless use of this latter step. To an extent the present paper can be construed as illustrating how badly things can go wrong if Nutter's warning goes unheeded.

[7]Here we are relaxing the "spec" approach a bit; but still it fits the philosophy for an omnithinker: no claims are being made regarding actual implementations, and the new sequence corresponds to technical features of the three standard formalisms, as will be pointed out later. Thus we have a kind of meta-example of the spec approach.

[8]Among philosophers, Dennett (1978) and Harman (1973, 1986) have studied this question in ways congenial to our approach.

internally consistent and deductively closed, and that is that. It is as if any statement whatever may count as a belief. For our purposes, this is insufficient; therefore we shall spend some time discussing this matter, and especially how it relates to the issues of consistency and tentativity in commonsense reasoning.

In everyday language, the word "believe" is used in several rather different ways. For instance, one might hear any of the following statements: "I believe you are right," "I believe Canada and Mexico are the only countries bordering the United States," "I believe this is the greatest country in the world," and "I believe gravity causes things to fall." These seem not to employ the same sense of the word "believe." Moreover, "two plus two equals four" can be regarded as a statement believed by whoever asserts it, and yet "I believe two plus two equals four" seems to convey a sense of less assurance than the bald assertion of the believed statement without the self-conscious attention to the fact that it is believed. An especially thorny aspect of this is that even when the statement of belief seems unambiguous, what is it that makes it true (or false)? That is, I may claim to believe x, but how do I mean this? That x is "in my head" seems both the obvious answer and yet completely misguided, for many things can be in my head without my believing them. That $2 + 2 = 5$ is surely in my head now as I write it, yet as something I *don't* believe. So it is a special "mark" of belief that makes certain things in the head beliefs. What "mark" is this, and what has it to do with reasoning?

We need then a definition of beliefs, so that we can make precise claims and attempt to defend them. Unfortunately, a precise definition will not be forthcoming here; it is a subject of considerable difficulty. However, a tentative but useful answer may be as follows: Certain things in the head (or within a reasoning system) may be *used in reasoning* as steps in drawing conclusions as to plans of action; let us call these *use-beliefs*. In effect, use-beliefs are simply potential steps in proofs of plans. (Note that these must be *genuine* steps, not convenient hypotheses, as in a natural deduction case argument, which later are discharged. That is, any such step must itself be a possible terminal step, i.e., a theorem.) Such entities would seem to form an interesting class of objects, relevant to the topic of commonsense reasoning. It is not offered as a matter of contention or empirical verification, but rather as an aspect of reasoning worth study. The main issue we wish to address then is the mutual consistency of use-beliefs in a commonsense reasoning system. For somewhat greater definiteness, we codify our "definition" below:

Definition

α is *(use-)believed* by agent g, if g is willing to use α as a step in reasoning when drawing conclusions as to plans of action.[9]

Whether an assertion A is to be a use-belief may depend on the context in which it is to be used. Thus in some contexts one and the same assertion "I believe X" may correspond to the presence of a belief X in our sense, and in another not, depending on the speaker's willingness to use X in planning and acting. Roughly speaking, the word "belief" will be used to refer to any strong notion that the agent is willing to trust, and does trust and use in planning and acting, "as if it were true." Now, this is not without its murky aspects. Many assertions can be sincerely doubted and sincerely taken as plausible at the same time (see Nutter (1982, 1983*a*,*b*) for a well-argued point of view on this). Instead of attempting to provide a foolproof analysis of beliefs by defining precisely the words "willing," "conclusion," "plan," etc. in the above definition, we will rely on examples.

Consider the following: That my car is still where I last parked it seems very likely, and I may well behave as if I regarded this to be the case, and yet I also recognize it is merely highly probable. Our proposal is to take such a statement to be a belief *if* I am prepared to use it as if it is true, and not otherwise. That is, if I regard it as highly probable but also hedge my bets by checking to see whether a bus will pass by in case the car is missing, then I do not have the belief that my car is where I last parked it; rather I may have the belief that *the probability is high* that my car is where I last parked it. However, if I ignore the possibility that my car may not be where I last parked it and *base my actions* on the assumption that it is still there, then even though I may admit that I am not certain where it is, I have the belief that it is where I last parked it. That is, we are defining the word "belief" in this manner (which incidentally seems consistent with one fairly common usage: one might very well say, "I realize my car may somehow have moved, but I nonetheless believe it has not").

Another example illustrating the difficulties in pinning down use-beliefs is the following (due to Michael Miller): Individual X believes (or so we wish to say) that smoking is dangerous to one's health, yet X does not give up smoking. Can we fairly say X is using that "belief" in planning and acting? I think that the answer is yes, but in a qualified sense. X will make use of Dangerous(smoking) as a fact to reason with, but this need not mean going along with (putting into action) the conclusion that X "should" give up smoking. (Compare Newell's principle of rationality (Newell 1981) in which "action P causes Q" and "Want Q" lead to "Do P.") What we need to do is to create (perhaps only as a thought experiment) a "neutral" or *ceteris paribus* ("other things being equal") situation in which X's actions can be supposed to be influenced only by the relevant "beliefs" we are testing. We are not in any way taking the view that X's behavior is determined only by a set of formulas in X's head (rather than by, say, stubbornness or competing concerns).

In other words, there may be many competing use-beliefs in one agent; this is not to be construed as all of them leading independently to given actions. Only in rarefied circumstances will this be the case. For instance, we may imagine a situation in which the belief that smoking is unhealthy would lead to a direct action in the given agent. For instance, if the agent's 12-year-old niece wishes to begin smoking, and if the agent is very concerned about her health, and if the agent lives thousands of miles away and would derive no comfort from another smoker in the family, and so on, *then* the given belief will lead to action to discourage the niece from smoking. Of course, the trick here is the "and so on." The definition then serves a purpose in being suggestive rather than definitive. What we need is a far deeper understanding of "what is in the head," perhaps something along the lines of Levesque's[10] notion of "vividness."

Are there situations where a more definitive kind of belief is available? Perhaps in the realm of scientific law, one can have beliefs that are held to be certain. It is worth exploring further examples. For instance, a belief in the actuality of gravity. As a physical theory, this is well supported, and yet most physicists

[9]An anonymous referee suggested the friendly amendment: "agent A believes that p in case, if A desires e, A is disposed to act in a way that will produce e given that p is true." I regard this as in the spirit I am aiming for; however, see below on *ceteris paribus* conditions.

[10]Levesque, H. 1985. Ninth International Joint Conference on Artificial Intelligence, Los Angeles, CA. Unpublished address.

will likely say that any theory of gravity is after all only an approximation that will almost surely be replaced by a better theory in the future, and indeed perhaps a theory in which gravity as such does not figure at all. Now, gravity may become a derived notion in such a future theory, but then it plays the role of naming a class of macroscopic phenomena such as "when I release a cup I have been holding, it will drop." But in what way is this a belief? Surely in the same sense as any everyday claim. That is, we expect the cup to drop, but do not regard it as absolutely beyond question. It may stick to our fingers, someone may catch it, a gust of wind may carry it upwards, and so on. If we try to eliminate "extraneous" factors such as these, we simply end up with the familiar qualification problem. It is not clear that any assertions of a general nature having practical consequences can be stated with certainty outside the realm of basic science where extraneous factors can be stipulated in full (at least relative to the theory one is using), and this lies far outside the realm of commonsense reasoning. Moreover, even in basic science, as we have illustrated, the so-called laws are usually taken to be tentative.

We can envision someone saying, "I definitely believe this cup will fall when I release my hold on it." And yet that same person will certainly grant our exceptional cases above, perhaps however protesting that these aren't what he or she had in mind. But that's just the point of the frame problem: We do not, and cannot, have in mind all the appropriate qualifications. We take the statement as asserted (the cup will fall when released) to be (or represent) the belief. Perhaps such thinking is more in the form of visual imagery than explicit statements to oneself, and perhaps visual imagery allows a certain loose notion of generic situations appropriate to commonsense reasoning. Be that as it may, people are often willing to assert boldly, even when questioned, that they believe such-and-such, and yet afterwards will agree that it is not so certain after all.

However, the "bare" unqualified assertions in Bel, by their definition as use-beliefs, are in fact significant as elements of thought, and it is also significant whether two or more elements of Bel conflict *in and of themselves*. To illustrate this, consider Reiter and Criscuolo's (1983) example of interacting defaults for Quakers and Republicans. Quakers (typically) are pacifists; Republicans (typically) are not. Nixon is a Quaker and a Republican. One might then tentatively conclude that Nixon is a pacifist, and also that he is not. Now, there is a sense in which this is quite reasonable: It does appear appropriate to consider that, on the one hand, Nixon might very well be a pacifist since he is a Quaker and, on the other hand, he might very well not be a pacifist since he is a Republican. That is, speculation about the origins and influences on this status regarding pacifism may be of interest for a given concern.

However, it will not do to entertain both that he is and that he is not a pacifist in one and the same planned sequence of actions. If one is using the assumption that Nixon is a pacifist, then one is not simultaneously willing to use the assumption that he is not a pacifist. That is, even though the expanded statements—that there is some evidence that Nixon might be a pacifist and some evidence that he might not—are not contradictory, nonetheless one would not be willing to assume, even for the sake of a very tentative kind of planning, that he both is and is not a pacifist. Moreover, if probabilities are used so that, let us suppose, Quakers have a 99% chance of being pacifists and Republicans a 90% chance of not being pacifists, then the recognition that these should not both be used separately to form tentative conclusions as to Nixon's pacifism depends on noting that such unexpanded conclusions—Nixon is a pacifist and Nixon is not a pacifist—indeed do conflict.

Consider again the released cup that may or may not fall. The speaker who claimed it would fall may enact a plan to cause the cup to fall, by releasing the cup. But if there is honey on the edge of the cup so that when released it remains stuck to the fingers, this plan will not be effective. However, when the honey is pointed out, thereby creating a belief that the cup will not fall when released, the direct contradiction with the earlier claim becomes important for the correct forming of a new plan. Now one must remove the honey or use more force in separating the cup from fingers, etc. However, if the original claim were to be qualified so that it accounted for the possible presence of honey, then the original plan of releasing the cup is not so easily accounted for. The point is that we do try to keep our unexpanded (use-)beliefs consistent, even though we may recognize that they are not strictly justified. A plan involves a package of tentative beliefs which are intended to be internally consistent, so that they can be enacted. Thus even if the planning agent may not firmly believe, say, the cup will definitely fall when released, still, his willingness to act as if he so believed appears to mandate his *not* believing, even tentatively, that the cup will also *rise* when released.

The point we are illustrating is that we usually do not tolerate direct[11] contradictions in the use-beliefs that enter into any one plan. If we decide to take a statement A as an assumption for purposes of a certain tentative line of reasoning, we ordinarily will not allow ourselves to also assume some other statement B that contradicts A in that same line of reasoning, even if we realize that both A and B are merely possibilities. Nonetheless, we shall argue below that contradictions do arise within our use-beliefs. Our previous arguments about the undesirability of contradictions among use-beliefs are intended to show that the presence of such a contradiction cannot be taken lightly. Once we have presented the form of contradiction we have in mind, we then will discuss its significance for formalizing commonsense reasoning.

The notion of use-beliefs has been described here at considerable length. This may have seemed overkill, especially on a topic that remains quite befuddled even as we come to the end of this section. Nonetheless, it has been a necessary exercise, for we wish to offset a possible objection to the analysis of default reasoning we will now pursue. In particular, we intend to analyze defaults in terms of beliefs. Now, it can be argued that the consequences of default reasoning, such as that Tweety can fly, are more properly regarded as tentative notions *that are not actual beliefs* (see Nutter (1982, 1983*a*)). However, the discussion of use-beliefs was intended to show that even tentative conclusions of this sort, if they are potentially *used* in any significant way in planning activity, should be treated much as if they were simply asserted flatly without qualification of tentativity.[12] That is, in particular, inconsistencies among such beliefs is a serious matter, more so than would be suggested by treating them as "shielded" by a "Tentative" modality. So we contend that, far from believing very little, commonsense reasoners will have very many use-beliefs.

3. A preliminary analysis of defaults

We now return to our earlier extended example of default

[11]More on this later.

[12]Much as if, but not wholly as if. This distinction will be brought out more later.

reasoning and analyze it more carefully. The expanded sequence of steps given in Sect. 1 was as follows:

(1) Bird(Tweety) [axiom]
(1*a*) Unknown(¬Flies(Tweety)) [oracle]
(1*b*) Unknown(¬Flies(x)) & Bird(x) .→ Tentative(Flies(x)) [default]
(1*c*) Tentative(Flies(x)) → Flies(x) [jump]
(2) Flies(Tweety) [consequence of 1, 1*a*, 1*b*, 1*c*]
(3) Ostrich(Tweety) [new axiom]
(3*a*) Ostrich(x) → ¬Flies(x) [axiom]
(4) ¬Flies(Tweety) [consequence of 3, 3*a*]
(4*a*) Flies(x) & ¬Flies(x)) .→. Suppress(Flies(x)) & Atypical(x) [fix]
(4*b*) Atypical(Tweety) [consequence of 2, 4, 4*a*]
(4*c*) Suppress(Flies(Tweety)) [consequence of 2, 4, 4*a*]

The fact that *in order to take advantage of the absence of information to the contrary, that absence must be recognized*[13] is codified in step (1*a*). This recognition in general is not decidable,[14] and so appeal is made to an outside source of wisdom, an "oracle," which tells us that a given proposition (e.g., Flies(Tweety)) is consistent with what we know. In the nonmonotonic logic (NML) of McDermott and Doyle (1980) and the default logic (DL) of Reiter (1980), such an oracle is explicitly represented, although in significantly different ways: NML presents an axiom including a modal operator M (for consistency), whereas DL uses M as a meta-symbol in a rule of inference. In McCarthy's circumscriptive logic (CL), a "weak" oracle appears in the form of the circumscriptive axiomatization itself, which is a source of both advantage and disadvantage: it is computationally more tractable and less prone than NML to suffer inconsistency of certain sorts (as we will see later), but also fails to recognize certain typical situations for the same reason.

Note that (1*b*) and (1*c*) are combined into a single default axiom or rule in NML and DL. We have drawn out the presumed underlying notions to focus attention on (1*c*) in particular—the *jump*. That is, up to that point, the reasoning is fairly clearcut; but at the jump, something that is only tentative is treated as if it were outright true. Herein is the source of the familiar unsoundness of nonmonotonic forms of reasoning. The agent whose reasoning is being stylized in the above sequence has jumped to a (firm) conclusion, albeit with plausible grounds to do so; but in the very jump is the possibility of error. This can be regarded as the point at which tentative conclusions are elevated into use-beliefs. However, the earlier discussion of use-beliefs indicates that even tentative beliefs (unelevated into "truths") can be use-beliefs.

In step (4*a*) recognition is made of the fact that an earlier conclusion has been contradicted and that the situation must be fixed. Actually, much more than what has been recorded in (4*a*) is necessary to thoroughly deal with the clash, but for now we will leave it as that.

We then arrive at the following preliminary characterization of default reasoning: it is a sequence of steps involving, in its most general form, oracles, jumps, and fixes. That is, it is error-prone reasoning owing to convenient but unsound guesses (jumps), in which therefore fixes are necessary to preserve (or re-establish) consistency, and which makes the mentioned guesses by means of appeals to undecidable properties (oracles). We will make use of this characterization in what follows. The general thrust of our line of argument will be that contradictions (such as between Flies and ¬Flies) generate the need for fixes, and that both then are necessary for the evolution of self-reflective reasoning. That is, we claim that a kind of temporary inconsistency pervades commonsense reasoning, and is one of its principal drivers.

[13]This is not to say that such recognition need be conscious, but merely that some mechanism or other must perform it.

[14]By *any* deterministic mechanism, of a logical stripe or not.

4. Israel's argument

Israel (1980) offers an argument for the inconsistency of commonsense reasoning. Recall that we use the intials "OT" to stand for "omnithinker," i.e., an idealized reasoning agent that is intended to be able to carry out an appropriate form of default reasoning and other desiderata as will be specified later. Israel begins by stating that OT will have some false beliefs and will have reason to believe that is the case, i.e., OT will believe what we shall designate as *Israel's sentence, IS*:

$$(\exists x)[\mathrm{Bel}(x) \,\&\, \mathrm{False}(x)]$$

where Bel refers in the intended interpretation to the very set of OT's beliefs, i.e., Bel(x) means that x is a belief of OT. (Note that *IS*, if it is to be a belief itself, requires that beliefs be representable as terms, e.g., quoted wffs, and that a certain amount of self-reference is then at least implicit in whatever language is used.)

Now, if *IS* is true in the intended interpretation, it follows that OT has a belief that is not true, namely, one of the *other* beliefs. On the other hand, if all the other beliefs are true in the intended interpretation, then *IS* is paradoxical in the sense that, in this interpretation, it is equivalent to its own denial. However, this does not force OT's belief set to be inconsistent, for the intended interpretation is not the only possible one. Moreover, *IS* may well be true there, i.e., other beliefs of OT may be false, and indeed this is the much more likely situation, and apparently the one Israel has in mind.

One might think (and Israel suggests) that even when *IS* is true, more is forthcoming, namely that since OT believes all OT's beliefs (i.e., OT believes them to be true) and yet also believes one of them not to be true, then OT believes contradictory statements and therefore has an inconsistent set of beliefs. However, for interesting reasons, this does not follow. The hitch is in the necessity of pinning down all OT's beliefs (or at least a suitable subset containing the supposedly false belief). That is, the phrases "all OT's beliefs" and "one of them not to be true" do not refer directly, in OT's asserting them, to the same things.

For example, suppose OT has three beliefs: α, β, and *IS*, and let us further suppose α is false. To follow the above suggestion for deriving a contradiction, we would like to argue as follows. OT believes one of α, β, and *IS* to be false: $\neg(\alpha \,\&\, \beta \,\&\, IS)$. Also OT believes α and β and *IS*, hence believes $(\alpha \,\&\, \beta \,\&\, IS)$. But these are contradictory. However, this argument makes exactly the oracle fallacy that many (including Israel) have inveighed against. For OT to conclude False(α & β & *IS*) from *IS* (i.e., from $(\exists x)[\mathrm{Bel}(x) \,\&\, \mathrm{False}(x)]$), OT must believe, in addition to α, β, and *IS*, that these are OT's only beliefs. But this would be another belief! Of course, a clever encoding of α might allow it to state that it itself along with β and *IS* are OT's only beliefs, thereby avoiding the trouble of an extra belief unaccounted for. (Alternatively, OT may mistakenly believe the aforementioned three to be its only beliefs.) But this does not resolve the difficulty at hand, for it is not plausible to argue in

general that OT will at any given moment have a belief such as this, unless a means is presented by which OT can deduce such a belief. This is exactly where oracles come into the picture. OT must know that it doesn't know (or believe) anything other than the three stated beliefs.

Now if OT refers to its beliefs by means of a term S for the set of these beliefs, then OT may have a belief such as $(\exists x)[x \in S \,\&\, \text{False}(x)]$ as well as $(\forall x)[x \in S \leftrightarrow \text{Bel}(x)]$. But this is not contradictory, for OT will presumably not believe $\text{Bel}(x) \rightarrow \text{True}(x)$, given that OT believes *IS*. For each belief x of OT, indeed OT believes x (to be true). But that is not the same as believing the conjunction of these beliefs to be true. This is a peculiar situation. OT indeed uses full deductive logic, but cannot prove the conjunction of its beliefs (axioms), not because of deductive limitations so much as descriptive power: OT has no name that is tied formally to its actual set of beliefs. If a superfancy brand of oracle is invoked to present such a name and the assertion that all elements named by that term are true, then indeed a contradiction follows. But there is no obvious argument that such an oracle is a part of commonsense reasoning.

Another way to state this is that getting OT's hands on its *set* of beliefs is not trivial, if this is to be done in a way that makes OT's term for that set correspond effectively to the elements (and no others) of that actual set.

(An interesting counterpoint is that the *negation* of *IS*, namely, $(\forall x)(\text{Bel}(x) \rightarrow \text{True}(x))$, when coupled with a relatively uncontroversial rule of inference (from x infer $\text{Bel}(x)$, i.e., OT may infer that it believes x, if it has already inferred x), often *does* produce inconsistency! Essentially, if T is a suitable first-order theory subsuming the mentioned rule, then *IS* is a *theorem* of T, quite the opposite of contradicting T. See Perlis (to appear) for details on this and related results.)

Nevertheless, we shall presently see that Israel's argument can be revised in such a way as to bear significantly on formalisms for default reasoning. To address this, we return now to our analysis of default reasoning.

5. Default reasoning and commonsense

The current breed of formal default reasoning tends to ignore the eventuality of errors cropping up in the course of reasoning, attention having focussed more on the semantics of getting the right initial default conclusion. But it is clear that if this can be done, then a mechanism is required to "undo" such a conclusion in the light of further evidence, as our example with Tweety indicates. Indeed, to deny information about errors to OT amounts to allowing OT the following kind of clumsy reasoning hardly suitable for an ideal reasoner:

> Tweety is a bird; so (perhaps in Reiter's or McCarthy's version) Tweety can fly. Why do I think Tweety can fly? I do not know. But Tweety turns out to be an Ostrich, so Tweety can't fly. Did I say Tweety could fly? I do not know why I said that. Do I think any bird can fly unless known otherwise? No, I do not think that.

While this may contain no inconsistency, it also seems not to be a very impressive instance of commonsense. We are not here trying to poke fun at proposals in the literature, but rather simply to illustrate how far indeed they are from the ideal of commonsense that apparently motivates their study. This will not be news, but it is nonetheless worthwhile to draw the boundaries to see where to go next. A slightly more commonsensical version is as follows:

> Tweety is a bird; so (because I do not know otherwise, perhaps in McDermott and Doyle's version) Tweety can fly: $\text{Bird}(x)$ & $\neg$ Known($\neg$ Flies(x)) $.\rightarrow \neg$ Flies(x), and also $\neg$Known($\neg$ Flies(Tweety)); but Tweety is an Ostrich, so Tweety can't fly. Did I say Tweety could fly? I do not know why I said that, for I believe birds fly if I do not know otherwise, but in Tweety's case I know otherwise. Did I say $\neg$Known($\neg$Flies(Tweety))? I wonder why, for it's not true.

A still smarter version, one that appears to deal with errors appropriately and keeps track of its reasoning over time, is

> Tweety is a bird; so (I may as well assume) Tweety can fly; but Tweety is an Ostrich and so cannot fly after all; my belief that Tweety could fly was false, and arose from my acceptance of a plausible hypothesis. Do all birds that may fly (as far as I know) in fact fly? No, that's just a convenient rule of thumb.

Note that here too a contradiction (between past and present) arises, but out of an explicit recognition that the default rule leads to other beliefs some of which are false. Which are the "real" facts? OT must be able to stand back from (some of) its beliefs to question their relative accuracy, temporarily suspending judgement on certain matters so they can be assessed, without thereby giving up other beliefs (such as general rules of reason) which may be needed to assess the ones in question.

Now the above scenarios strongly suggest that for OT to be capable of appropriate commonsense reasoning, it must be able to reflect on its past errors, indeed, on its potential future errors. This observation will form the basis of our next section, in which we consider "Socratic" reasoners, i.e., ones that know something of their own limitations, and in particular the fallibility of their use of defaults.

6. Israel's argument revisited

There is a way to make Israel's argument good after all, by reformulating *IS* into a version, say, *IS'*, to make it refer to a particular proscribed set, indeed a finite one that can be listed. It is not essential at all for *IS'* to refer to all beliefs of OT, nor even to itself. It is sufficient that *IS'* refer to a finite set S of beliefs of OT that OT can explicitly conjoin into a single formula; this will then produce a contradiction if OT has the full deductive power of logic and if *IS'* is a belief of OT that states that at least one of the elements of S is false. Now, what reason can be given for OT having such a belief as *IS'*? Why would an intelligent reasoner such as OT concede that some subset S of its beliefs holds an error, especially since it is not just *some* subset, and not just a finite one, but one that can be explicitly divulged.

Our answer to this is that it is precisely default reasoning, i.e., reasoning by guess, uncertainty, or jumps that makes such a concession inevitable. For OT, to be truly intelligent, must realize that its default beliefs are just that: error-prone and therefore at times just plain wrong. Indeed, the very necessity of making fixes blatantly exposes the error-prone quality of default reasoning. Now we can get an explicit formal contradiction, for we can argue that *IS'* will reasonably be believed by OT. All OT needs is a handle on its own past, e.g., that it has judged many instances of a default to be positive (such as 1000 birds to fly).

For the purpose of the following definitions, we consider OT to be a "reasoning system," i.e., some version of a formal theory (actually, a sequence of such theories) that varies over time as it interacts with new information. Thus in our earlier example, at the point at which it is learned that Tweety is an ostrich, the

reasoning system is considered to have evolved into another formal state. However, as will be seen below, time and the elements of defaults (oracles, jumps, and fixes) serve only a motivational role, not needed for the formal treatment. We work then within an appropriate first-order theory, supplemented with names for wffs to allow quotation and unquotation as in Perlis (1985).

Definition

A reasoning system OT is *quandaried* (at a given time) if it has an axiom (belief) that says not all of an explicitly specified finite set of its beliefs are true.

Theorem 1

No quandaried reasoner can be consistent.

Proof

Let OT be a quandaried reasoner, for which the explicitly specified set of beliefs is $\{B_1, \ldots, B_n\}$, and having in addition the axiom (belief) $\neg(B_1 \ \& \ \ldots \ \& \ B_n)$. The result follows immediately.

Now while theorem 1 may seem trivial, it does hold an interesting lesson. For any reasoner that has perfect recall of its past reasoning will be able to explicitly specify its past default conclusions, and in particular those that have not been revoked (fixed). If it then also believes on general principles that one or more of *these* beliefs is false, it will be quandaried and hence inconsistent. We presently codify this in further definitions and a theorem. Note, however, that theorem 1 does not necessarily spell despair for quandaried reasoners. For the spirit of the Tweety example, which has motivated much of our discussion, is precisely that an inconsistency giving rise to a fix that restores consistency. That is, it is to be expected that OT may fluctuate between quandaried and nonquandaried states, as it finds and corrects its errors.

Definition

OT is *recollective* if at any time t it contains the belief

$$(\forall x)(\mathrm{Dflt}(x) \leftrightarrow x = b_1 \ v \ldots v \ x = b_n)$$

where $b_1, \ldots, b_n$ are (names of) all the beliefs $B_1, \ldots, B_n$ derived by default prior to t (i.e., $b_i =$ "B_i"), and if for each i either OT retains the belief B_i as well as $\mathrm{Bel}(b_i)$, or else B_i has been revoked (e.g., by a fix) and then it contains the belief $\neg\,\mathrm{Bel}(b_i)$.

Definition

OT is *Socratic* if it has the belief $(\exists x)(\mathrm{Dflt}(x) \ \& \ \mathrm{Bel}(x) \ \& \ \mathrm{False}(x))$ as well as the beliefs $\mathrm{False}(\text{“}\alpha\text{”}) \rightarrow \neg\,\alpha$ for all wffs α.

Theorem 2

No recollective Socratic reasoner can be consistent.

Proof

Any such reasoner OT will be quandaried and so by theorem 1 will be inconsistent. To see that OT will indeed be quandaried, simply observe that in being Socratic and recollective, OT will believe

(i) $(\exists x)((x = b_1 \ v \ldots v \ x = b_n) \ \& \ \mathrm{Bel}(x) \ \& \ \mathrm{False}(x))$

But also (in being recollective) OT will believe either

(ii) $B_i \ \& \ \mathrm{Bel}(b_i)$

or

(iii) $\neg\,\mathrm{Bel}(b_i)$

for each b_i. Now consider those b_i such that $\mathrm{Bel}(b_i)$ is believed (i.e., is a theorem of OT), say, $b_{i_1}, \ldots, b_{i_k}$. Since OT believes $\neg\,\mathrm{Bel}(b_i)$ for the *other* beliefs among $b_1, \ldots, b_n$, then by (i) OT believes one of $b_{i_1}, \ldots, b_{i_k}$ to be false, yet each is believed; so OT is quandaried. (It is also not hard to form a direct contradiction without exploiting the notion of a quandary; we have pursued this route simply to illustrate the application of Israel's (modified) argument.)

As an example, OT may believe a rule such as that "typically birds can fly," and operationalize it with a (second) rule such as that given $\mathrm{Bird}(x)$ and if $\mathrm{Flies}(x)$ is consistent with OT's beliefs, then $\mathrm{Flies}(x)$ is true. But OT will also believe (since it is smart enough to know what defaults are about) that this very procedure is error-prone, and sometimes $\mathrm{Flies}(x)$ will be consistent with its beliefs and yet be false. Indeed, it will reasonably believe that one of its *past* (and yet still believed) default conclusions is such an exception, and these it can enumerate (suppose it has reasoned about 1000 birds) and yet it will also believe of each of them that it is true! So OT is inconsistent. Of course, the very realization of the clash (which OT also should be capable of noticing) should generate a fix which calls the separate default conclusions into question, perhaps to be relegated again to their more accurate status of tentativity. We will see in the next section how these results relate to the standard default formalisms in the literature.

It is necessary here to address the possible objection that OT will not be able to enumerate its past defaults and so will refer to them only generically, defusing the contradiction as we did with Israel's original argument. However, it is not at all unreasonable to suppose that OT keeps a list of its defaults, especially in certain settings. For instance, if OT works as a zookeeper and keeps a written record of the animals there, 1000 North American (i.e., "flying") birds may have been recorded by OT as in good health (and so able to fly), and OT may continue to defend these judgements even while granting that some of them will be errors. More will be said on this in the following section. (Readers may recognize this as a version of the Paradoxes of the Preface or of the Lottery.[15] See Stalnaker (1984).)

7. Analysis of three formalisms

We are now in a position to present rather striking examples of the situation that has been dealt with in the earlier sections. We will show that major weakenings occur when the standard "epistemological adequacy" approaches to commonsense reasoning are combined with a recollective *or* Socratic treatment of use-beliefs. We will examine McCarthy's circumscription (CL), McDermott and Doyle's nonmonotonic logic (NML), and Reiter's default logic (DL). Recall (theorem 2) that any Socratic *and* recollective default reasoner is inconsistent. This of course applies as well to CL, NML, and DL; that is, if any of these is endowed with Socratic and recollective powers, it will become inconsistent. This already is unfortunate, in that it seems to suggest that quite a different sort of formalism will be required to handle defaults "realistically." But even more damaging observations can be made, namely, relaxing in various ways the Socratic and recollective hypotheses still produces undesired results in these formalisms.

We begin by reviewing the extent to which our three default "keys" (oracles, jumps, and fixes) come into these treatments. The case of circumscription, or CL, is slightly complicated by the fact that the predicate "Unknown" (that is, Unknown-($\neg\,\mathrm{Flies}(x)$) in our example) is not explicit, and indeed, CL does

[15] McDermott (1982) mentions this paradox as one giving trouble for monotonic logics; here we see that it is also problematic for nonmonotonic logics.

not quite test fully for whether Flies(x) is already entailed by the given axioms, but rather uses a substitute (known as the method of inner models). This has the advantage of being semidecidable (i.e., the oracle is actually represented somewhat algorithmically, in the form of a second-order axiom schema), though it has the disadvantage of being incomplete (Davis 1980; Perlis and Minker 1986). Thus steps $1a$, $1b$, and $1c$ are implicit in CL, whereas in NML and DL step $1a$ is implicit (with an ineffective oracle used to supply full consistency information for Unknown) and steps $1b$ and $1c$ are combined into a single step (an axiom of NML and in inference rule of DL) stating directly (in the case of our example) that if Tweety is a bird and if it is consistent to assume Tweety flies, then Tweety does fly. In other respects, however, the three approaches are much the same. Steps $4a$, $4b$, and $4c$ are simply not present in any form in any of them, since once Ostrich(Tweety) is added as a new axiom, the previous axiomatization is no longer under consideration and it and its conclusions (e.g., Flies(Tweety)) are ignored.

It is clear then that to address the issues urged here, these approaches must be supplemented, and in particular with "histories" of their conclusions. More generally, far greater explicitness of world knowledge and of their own processes is required. But we have seen that when such information is allowed in a default reasoner, it has a high chance of becoming inconsistent. In particular, if it can represent the fact that it is using default rules, and that some of its default conclusions will therefore be erroneous, and if it can recall its default conclusions, then it is recollective and Socratic and so falls prey to theorem 2 and will be inconsistent. However, the particular features of CL, NML, and DL are such that simpler means exist to derive implausible conclusions within an intuitively commonsense framework. This is most easily illustrated in the case of NML because, of the three formalisms mentioned already, it alone has enough apparatus present to express the required concept of default fallibility directly.[16]

In NML, the Tweety example might be handled as follows. We could specify the axiom

$$(\forall x)((\text{Bird}(x) \,\&\, \text{MFlies}(x)) .\!\rightarrow \text{Flies}(x))$$

where M is a modal operator interpreted as meaning that the wff following it is consistent with the axioms of the formalism itself. Since this involves an apparent circularity, McDermott and Doyle go to some lengths to specify a semantics for M. However, for our purposes, it is not essential to follow them in such details; we can form a "Socratic" version of *IS'* easily for this case:

$$(\exists x)(\text{Bird}(x) \,\&\, \text{MFlies}(x) \,\&\, \neg\text{Flies}(x))$$

This at least partly expresses that, for some bird, NML has the needed hypotheses to infer by default that the bird flies, and yet it will not fly. One could hardly ask for a more direct statement of the fallibility of defaults. Yet now NML is in trouble; it will immediately find the following direct contradiction:

$$(\exists x)(\text{Flies}(x) \,\&\, \neg\text{Flies}(x))$$

Note that here we do not need the recollective hypothesis since a Socratic[17] extension of NML is already inconsistent. (Nutter (1982) and Moore (1983) have pointed out that the formalization of NML seems to commit an error of representation vis-a-vis intuitive semantics, and it is this in effect that we are exploiting here.)

Even though, as we have seen above, in NML it is possible to express a Socratic-type axiom that some *default* conclusion has gone awry, thereby producing inconsistency, one can make an even (apparently) weaker assumption and still achieve distressing results. In all three formalisms, a simple additional "counter-example" axiom seems to undermine the ability of the reasoning system to make appropriate default conclusions consistently, surrendering the full Socratic condition but (for DL and NML) employing the recollective condition.

For instance, given the zookeeper axiom

$$(\exists x)(\text{Bird}(x) \,\&\, \neg\text{Flies}(x) \,\&\, x = b_1 \vee \ldots \vee x = b_{1000})$$

Reiter's DL "sanctions" the conclusion Flies(b_i) for each bird b_i separately. Now a problem arises: How are we to interpret these sanctions? Reiter apparently intends that any wffs entailed by *all* default extensions are to be treated as default conclusions, and that others *may* be so treated if we are careful not to mix such from distinct extensions. In any case, we are stuck, because we need the zookeeper to make a sequence of such conclusions that do *not* fall into any one extension. That is, DL augmented by sufficient axioms to specify all the (finitely many) birds in the zoo might conclude of each one by one that (it is "ok" to suppose that) it flies, up until the last one, and since it also has the belief that one of them does not, it will be forced to conclude of the last that *it* is the culprit that does not fly. Therefore, DL (so construed) is prey to the "paradox of the zookeeper" in that, although avoiding inconsistency, its ability to derive the intuitive default conclusions is compromised.[18] The same arguments apply to NML and counter-example axioms. Alternatively, if the time sequence is finessed, these formal specifications of default reasoning can be viewed as producing the conclusion that precisely one bird of the 1000 does not fly, without committing themselves to even a single default conclusion about any individual bird. This, however, amounts to avoiding drawing any atomic defaults, largely defeating the prime motivation for such reasoning.[19] This will arise even more dramatically when we examine CL below.

Here our earlier treatment of use-beliefs comes in handy. For we can argue that the zookeeper "really" believes that each separate bird at the zoo can fly, that he is highly unwilling to leave any of their cage doors open, and that he is also unwilling

[16]And this is its downfall regarding our present discussion. In effect, NML makes an axiom out of DL's rule. Thus DL's ability to refuse to believe the rule as a truth is not available to NML. Of course, this is indulging in an introspectivist fantasy, for neither formalism represents reasons for its conclusions. But this fantasy does, I think, accurately pinpoint the critical distinction that allows DL to survive the threat of inconsistency in the present example.

[17]We use the term "Socratic" somewhat loosely here, since it is not precisely the same as the formal definition given earlier. However, intuitively it still expresses the idea of self-error.

[18]That is, the zookeeper concludes of *no* bird that it does not fly; only the existence of such is believed. I grant that, to get DL, NML, and CL to take the job of zookeeper, I am stretching them in unintended ways and making arbitrary choices in the process. But that is the point: We must devise formalisms that do lend themselves to introspection; and when we do, there will be difficulties of the sorts described here.

[19]It also seems related to Reiter's suggestion that DL has a representation of the notion of "few" or "many." However, the actual inferences sanctioned by DL (or by NML or CL, for that matter) seem to miss much of the import of these terms, in that a strict minimum is determined. For example, that most birds fly, when represented as a default, leads to the conclusion that *all* birds fly if no counter examples are known, and that *all but* known counter examples fly in other cases. But there is an intended indefiniteness in the words "few" and "many." See our discussion of CL below.

to call any one of them to the attention of the zoo veterinarian. Yet, he is also very concerned at the veterinarian's failure to arrive for work at the usual hour, because the zookeeper also believes that *some* (unspecified) birds in the zoo *are* ill (and unable to fly).

There is an intuitive appeal to this, in that a zookeeper might very well defend each separate conclusion of the form Flies(b_i), and yet not agree to the statement that all the birds fly. Indeed, the zookeeper has *use-beliefs* Flies(b_i) for each i as well as the use-belief $(\exists x)(\neg \text{Flies}(x) \ \& \ x = b_1 \ v \ldots v \ x = b_{1000})$. The problem is that whereas the zookeeper may refuse to apply an inference rule to form the conjunction of given beliefs (recognizing, in effect, that there is a contradiction afoot and that his beliefs are tentative), DL and NML are formal extensions of first-order logic and will have such conjunctions as theorems, with no means to control the usual disastrous consequences of this in logically closed formalisms. That is, zookeepers and other commonsensical beings are perhaps not obedient to slavish rules of formal logic regarding their use-beliefs.[20]

One might seek to alter the logic as a way around this, for instance by not allowing arbitrary conjunctions of theorems. Such alternatives have been raised in connection with the lottery and preface paradoxes (Stalnaker 1984). However, this does not affect our conclusion that the beliefs of any agent fully in the zookeeper's shoes would in fact be mutually (but not directly) contradictory, whether or not there are mechanisms in place to keep the contradiction from being deduced. Moreover, an intelligent agent should be able to recognize the contradiction and conclude, perhaps, that its conclusions of the form Flies(x) were after all only tentative (that is, undo the "jump"). For this, a record must be kept of the origin of conclusions, a point utilized in Doyle (1979) and emphasized in Nutter (1983*a*).

It is worth contrasting the zookeeper scenario with another, the "detective" scenario. Here there are, say, 10 suspects in a murder case, each of whom has an alibi and is apparently a very nice person. Yet instead of concluding separately of each that he or she is (tentatively) innocent, our detective tentatively suspects each (separately) of being guility. But if there were 1000 or more suspects (e.g., if relatively little at all existed in the form of clues, so that *anyone* may have been the murderer), then it is no longer reasonable to tentatively treat each individual as guilty. That is, in the case of 10 suspects, it may be life-preserving to be wary of all 10, but in the case of 1000, it surely is counter productive. This seems to say that raw numbers do (should) affect the course of default reasoning, a matter we will not pursue further here.

The case of CL is still more interesting. Here, the defaults are not as explicitly represented as in NML (or even DL).[21] So, following the cue of our additional axiom above, we say simply that there is a bird that does not fly, as well as circumscribing nonflying birds. Thus

$$\{\text{Bird(Tweety)}, (\exists x)(\text{Bird}(x) \ \& \ \neg \text{Flies}(x))\}$$

when circumscribed with respect to $\neg$Flies (letting Flies be a "variable" circumscriptive predicate), instead of providing the expected and usual (when the second axiom is not present) conclusion that Tweety flies, allows us no conclusion at all about Tweety not already contained in the axioms before circumscribing.[22] That is, from Bird(Tweety) *alone* circumscription of $\neg$Flies produces Flies(Tweety), as desired. Yet with the additional axiom present, this no longer is the case. This is easily seen, for the above axiom set has a minimal model in which Tweety is precisely the claimed exceptional bird. Here we do not even need the recollective condition.

We wish to regard Tweety as a typical bird, since nothing else is known explicitly about Tweety; however, the additional axiom raises the possibility that Tweety may not fly, i.e., Tweety may be the intransigent nonflying bird that is asserted to exist. The result is that *no* birds will be shown (even tentatively) to fly by circumscribing in such a conext. Clearly this is not what we wish of a default reasoner. In terms of our analysis of default reasoning, CL does not perform step 1*a*; the circumscriptive schema (oracle), when faced with a weak Socratic axiom, no longer recognizes the "typical" case. The result claimed above is formalized below.

Theorem 3

The set

$$T = \{P(c), (\exists x)(P(x) \ \& \ \neg Q(x))\}$$

when circumscribed with respect to $\neg Q$, does not have $Q(c)$ as theorem, even in formula circumscription with P and Q allowed as variable predicates.

Proof

By the soundness theorem for circumscription[23] (see McCarthy (1980) and Perlis and Minker (1986)), if $Q(c)$ were a theorem of Circum(T, $\neg Q$), then $Q(c)$ would hold in all minimal models of T (with respect to $\neg Q$). But this is not so. There are minimal models of T in which $\neg Q(c)$ holds, namely, ones in which c is the only P-entity. (Intuitively, c = Tweety is the only bird (P) that does not fly (Q).)

One might try to "disconnect" Tweety from the existentially asserted nonflying bird, for instance, by Skolemizing the additional axiom as

$$\text{Bird}(c) \ \& \ \neg \text{Flies}(c)$$

However, this will not work either. We still cannot prove Flies(Tweety) by circumscription, unless we adopt the further axiom that Tweety $\neq c$. But to do this amounts to begging the question, i.e., to assuming we already know (before we circumscribe) that Tweety is not to be exceptional in regard to

[20]We might say the zookeeper is *quasi-consistent* (and *quasi-inconsistent*), that is, his belief set B *entails* a contradiction X & $\neg$X, but does not *contain* one directly (X and $\neg$ X are not both elements of B). It follows that such an agent cannot be logically omniscient, but in a way that is no weakness at all. The entailed contradiction need not be *missed* out of logical ignorance, but rather can be deliberately rejected *on the basis* of having been duly (and logically) noted and judged impossible and attributed to an error.

[21]Although *formula* circumscription does provide at least part of a mechanism for expressing within the logic the fact of self-error, and if this is teased out by means of suitable metalogical devices, then in close analogy with the following treatment commonsense is compromised.

[22]That is, about Tweety in isolation. As seen above with NML and DL, shotgun results about the whole set of birds may be derivable, such as that there is only one nonflying bird, but no conclusion specific to any particular bird follows. Indeed, the easily proved circumscriptive result that there is only one bird that does not fly (also derivable in NML and DL when interpreted as in our earlier discussion in the recollective case for zoo birds) runs counter to intuition: the existence of a nonflying bird suggests "few," not "only one." It would be much more satisfactory if there were a way to remain noncommittal on the exact number.

[23]Etherington, D. 1984. Week on logic and AI. University of Maryland. Private communication.

flying. Moreover, we can then simply consider the wff $\neg$Flies(x) & $x \neq c$ instead, and assert (reasonably) that some bird satisfies *this*. For if c were the only nonflying bird, then we would not need defaults in the first place. The whole point is that even among those birds that seem typical as far as we can tell, still there lurk exceptions.

This can be seen in a more dramatic form by postulating that $b_1, \ldots, b_n$ are *all* the birds in the world (where n is some large known integer, say, 100 billion, and distinct b_i's may or may not represent distinct birds). Then the axiom $(\exists x)$(Bird(x) & $\neg$Flies(x)) cannot be Skolemized with a constant that is also assumed to have a distinct reference from every b_i. Note that this is similar to Reiter's unique names hypothesis (1980) that likewise is not handled directly by circumscription. We suggest that a solution to one of these problems may harbor a solution to the other. Note, however, that even if names are introduced into CL in such a way that one can circumscriptively prove $\neg$Flies(Tweety) (i.e., so that jumps are reinstated), then CL immediately falls prey to inconsistency if it is also recollective. (See Etherington *et al.* (1985) for the crucial role of existential quantifiers in circumscriptive consistency.)

What we have found, then, regarding "realistic" default reasoning and three standard formalisms in the literature, can be represented in the following table:

	Formalism		
Hypotheses	CL	DL	NML
Socratic	compromised	—	inconsistent
Recollective	—	—	—
Socratic–recollective	inconsistent	inconsistent	inconsistent
Counter example	compromised	—	—
Counter example–recollective	compromised	compromised	compromised

Here "compromised" means that the formalism in question will not produce the intuitively correct commonsense default conclusions, and "—" means that no apparent difficulties arise. So we see that any of the three formalisms can be made recollective without upsetting the intended usage. On the other hand, the Socratic or counter-example conditions, which are what express the *tentativity* (i.e., the default-hood) of defaults, tend to spell trouble. It is of interest that Reiter's version, DL, comes out "best" of the three: it suffers the least affront to its default integrity as a specification for an omnithinker.

Finally, in no case can the "ideal" of a Socratic recollective default reasoner be achieved within the framework of the epistemological adequacy approach, since that approach is based on the assumption of a consistent, logically closed axiomatization. This is of course relative to our definition of use-beliefs; that is, the inconsistency may lie hidden inside tentativity predicates, as Nutter (1983*b*) urges. But the three formalisms discussed in this section all employ jumps, i.e., they baldly assert their default conclusions, and therefore the implicit use-belief inconsistency that will arise when they are endowed with Socratic-recollective features will become an explicit logical inconsistency.

8. Conclusions

We need formalisms adequate to the task of capturing "introspective" default reasoning. This is essential to performing certain kinds of fixes. Furthermore, the latter often cannot be done at all without sacrificing consistency in favour of a kind of quasi-consistency.

The thrust of our remarks has been that desiderata underlying commonsense reasoning simply are inconsistent, and that we now must devise and study systems having such characteristics. Specifically, the "jump" phenomenon, so central to most work on default reasoning (and, notably, attacked by Nutter (1982)), will not withstand simultaneous admission of fallibility implicit in a "fix." On the other hand, dealing with inconsistent formalisms seems to force us toward deeper analysis of processes of memory, inference, and focus over time. These are the topics of work in progress (Drapkin *et al.* 1986; Drapkin and Perlis 1986). Nutter (1983*b*) prefers to avoid jumps and seeks other means of utilizing default conclusions such as relevance logic. The extent to which the term "logic" is appropriate at all for such an undertaking is also a matter of debate (see Doyle (1983, 1985) and Harman (1986)). It will be interesting to see whether any of these approaches bear fruit.

Acknowledgements

I wish to thank James Allen, Jennifer Drapkin, Jerry Feldman, Rosalie Hall, Jim Hendler, Hector Levesque, Vladimir Lifschitz, Ron Loui, Michael Miller, Jack Minker, Dana Nau, Rich Pelavin, Jim des Rivieres, John Schlipf, and Jay Weber for useful discussion on the topic of this paper. Special thanks to Pat Hayes and Ray Reiter, who challenged me to write down and clarify my spoken claims.

This research has been supported in part by the U.S. Army Research Office (DAAG29-85-K-0177) and the Martin Marietta Corporation.

DAVIS, M. 1980. The mathematics of non-monotonic reasoning. Artificial Intelligence, **13**, pp. 73–80.

DENNETT, D. 1978. Brainstorms. Bradford Books, Montgomery, VT, pp. 300–309.

DOYLE, J. 1979. A truth maintenance system. Artificial Intelligence, **12**, pp. 41–72.

——— 1983. A society of mind. Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany, pp. 309–313.

——— 1985. Reasoned assumptions and Pareto optimality. Proceedings of the Ninth International Joint Conference on Artificial Intelligence, Los Angeles, CA, pp. 87–90.

DRAPKIN, J., and PERLIS, D. 1986. Step-logics: an alternative approach to limited reasoning. Proceedings, Seventh European Conference on Artificial Intelligence, Brighton, England.

DRAPKIN, J., MILLER, M., and PERLIS, D. 1986. A memory model for real-time default reasoning. University of Maryland, College Park, MD, Technical Report.

ETHERINGTON, D., MERCER, R., and REITER, R. 1985. On the adequacy of predicate circumscription for closed-world reasoning. Computational Intelligence, **1**, pp. 11–15.

HARMAN, G. 1973. Thought. Princeton University Press, Princeton, NJ.

——— 1986. Change in view. MIT Press, Cambridge, MA.

ISRAEL, D. 1980. What's wrong with non-monotonic logic? Proceedings of the First National Conference on Artificial Intelligence, Standford, CA, pp. 99–101.

KONOLIGE, K. 1984. Belief and incompleteness. SRI International, Menlo Park, CA, Technical Note 319.

MCCARTHY, J. 1980. Circumscription—a form of non-monotonic reasoning. Artificial Intelligence, **13**, pp. 27–39.

——— 1986. Applications of circumscription to formalizing commonsense knowledge. Artificial Intelligence, **28**, pp. 89–116.

McDERMOTT, D. 1982. Non-monotonic logic. II. Journal of the Association for Computing Machinery, **29**, pp. 33–57.

McDERMOTT, J., and DOYLE, J. 1980. Non-monotonic logic I. Artificial Intelligence, **13**, pp. 41–72.

MOORE, R. 1977. Reasoning about knowledge and action. Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, MA, pp. 223–227.

——— 1983. Semantical considerations on non-monotonic logic. Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany, pp. 272–279.

NEWELL, A. 1981. The knowledge level. AI Magazine, **2**, pp. 1–20.

NUTTER, J. T. 1982. Defaults revisited, or "Tell me if you're guessing". Proceedings, Fourth Conference of the Cognitive Science Society, Ann Arbor, MI, pp. 67–69.

——— 1983*a*. What else is wrong with non-monotonic logic? Proceedings, Fifth Conference of the Cognitive Science Society, Rochester, NY.

——— 1983*b*. Default reasoning using monotonic logic: a modest proposal. Proceedings of the International Joint Conference on Artificial Intelligence, Washington, D.C., pp. 297–300.

PERLIS, D. 1981. Language, computation, and reality. Ph.D. thesis, University of Rochester, Rochester, NY.

——— 1985. Languages with self-reference I: foundations. Artificial Intelligence, **25**, pp. 301–322.

——— 1986. Languages with self-reference II: knowledge, belief, and modality. Artificial Intelligence, to appear.

PERLIS, D., and MINKER, J. 1986. Completeness results for circumscription. Artificial Intelligence, **28**, pp. 29–42.

REITER, R. 1980. A logic for default reasoning. Artificial Intelligence, **13**, pp. 81–132.

REITER, R., and CRISCUOLO, G. 1983. Some representational issues in default reasoning. International Journal of Computers and Mathematics (Special issue on computational linguistics), **9**(1), pp. 15–27.

STALNAKER, R. 1984. Inquiry. MIT Press, Cambridge, MA, pp. 90–92.