In *Computational Intelligence*, Volume 13, Number 2, 1997, pp. 165-187

WHAT SHOULD DEFAULT REASONING BE, BY DEFAULT?

FRANCIS JEFFRY PELLETIER
Department of Computing Science and Department of Philosophy, University of Alberta, Edmonton

RENÉE ELIO
Department of Computing Science , University of Alberta, Edmonton

This is a position paper concerning the role of empirical studies of human default reasoning in the formalization of AI theories of default reasoning. We note that AI motivates its theoretical enterprise by reference to human skill at default reasoning, but that the actual research does not make any use of this sort of information and instead relies on intuitions of individual investigators. We discuss two reasons theorists might not consider human performance relevant to formalizing default reasoning: (a) that intuitions are sufficient to describe a model, and (b) that human performance in this arena is irrelevant to a competence model of the phenomenon. We provide arguments against both these reasons. We then bring forward three further considerations against the use of intuitions in this arena: (a) it leads to an unawareness of predicate ambiguity, (b) it presumes an understanding of ordinary language statements of typicality, and (c) it is similar to discredited views in other fields. We advocate empirical investigation of the range of human phenomena that intuitively embody default reasoning. Gathering such information would provide data with which to generate formal default theories and against which to test the claims of proposed theories. Our position is that such data are the very phenomena that default theories are supposed to explain.

Keywords: foundations, default reasoning, cognitive science, nonmonotonic reasoning, psychologism

1. INTRODUCTION

Our interest in this paper is to consider how the AI field defines the phenomenon of default reasoning is, and what it presents as the underlying *point* or *purpose* of constructing formal accounts of it. We were motivated to reconsider these issues at this high level in part to ground our own research agenda in the concerns of the field, and in part to find a unifying framework with which to understand the background motivations of AI research into default reasoning. We believe that attention to these questions and their implications can help direct future formal research into the logical systems of, and the practical implementations for, default reasoning. We also believe that attention to the answers can set an interesting and practical research agenda. And finally, we believe that the results of this research agenda can have important relationships with other areas of research that are themselves concerned with default reasoning in one form or another.

We address two main topics. The first topic outlines the dangers of the methodology that seems to underlie the development and testing of default reasoning theories. The second topic is presented as part of a prescription for the first, namely, that considering how *people* actually do default reasoning is an important and necessary grounding for the entire enterprise of formalizing default reasoning. We expect that most researchers will be sympathetic to the litany of concerns that we present as the first topic; indeed, some may argue that these methodological concerns are

Address correspondence to F. J. Pelletier at the Department of Computing Science, University of Alberta, Edmonton, Alberta, T6G 2H1, Canada. Email: jeffp@cs.ualberta.ca

obvious and well-appreciated. By the same token, we expect that there may be less acceptance of our prescription to ground formalizing default reasoning in human default reasoning. We give due consideration to some of these arguments concerning such a prescription in the final sections of this article.

2. BACKGROUND AND SCOPE

Broadly speaking, our topic of concern is drawing conclusions when it is known that more information might be relevant, and therefore the inference is corrigible. Our topic also covers drawing conclusions when there is conflicting information that gives us reason to want to draw differing conclusions, depending on how the available information is used. These issues have been long studied in other fields.[1] Default reasoning captured the attention of the AI field in roughly the late 70's, when the Knowledge Representation (KR) community argued that this sort of reasoning was outside the scope of classical deductive logic, and yet was central to the description and construction of artificially intelligent agents. Since that time, much work of a formal nature has been done developing formalisms and mechanisms that were to give *some* account of these sorts of inference. Different formalisms emphasized different aspects of these inferences and a plethora of names have evolved to designate certain ways of looking at the enterprise. This makes it somewhat difficult to employ terminology which is neutral with respect to all the theories involved, yet this is what we wish to do because we wish to talk about the phenomenon in its broadest sense that these theories attempt to systematize. In this paper, we therefore settle upon the term 'default reasoning' as the general title of the area under investigation. We might have used the term 'nonmonotonic' rather than 'default', since the term 'monotonic' has a fixed meaning in classical logic: that the set of conclusions drawn from a set of premises never shrinks as more premises are added. We have decided against adopting this term because it suggests that a formal system, perhaps classical logic with some added component for defaults, will be the way to account for default reasoning. And not all of the theories we wish to include in our general discussion believe this.[2] We also understand the use of the term 'logic' to presuppose giving a theoretical account of the phenomena by means

---

[1] An inventory of the other areas in which default reasoning has been studied would include, at least, ethics, legal reasoning, causal reasoning, medical diagnosis, and judgment and decision making. Ethics in particular has had a long history of investigation into default reasoning. Ethical "laws" are commonly thought to confer only *prima facie* obligations, they are always accompanied by *ceteris paribus* conditions; similarly we only have *prima facie* rights. These obligations, duties, and rights are real, yet they are defeasible: further factual information might come in to demonstrate that we do not have the obligation, duty or right in some specific circumstance. And as ethicists have long noted, it is impossible to specify completely all the possible *ceteris paribus* clauses for any of these types of "law" in order to thereby make the default rule become an absolute law. For if one could specify all the factually relevant possible exceptions and thereby make the ethical law become an "empirically descriptive" statement, one would in effect have derived an 'ought' statement from an 'is' statement–i.e., we would be able to derive an ethical statement from what is in fact happening. And this is something that no one thinks is possible.

[2] Poole (1988) for example, discusses the Theorist framework as one in which default conclusions are hypotheses to be explained, and the semantics of classical logic can be used to construct these explanations.

of a formal system employing rules of inference.[3] Although all the theorists we consider in this paper agree that a formal system is to be used for this purpose, it is well to remember that there are other, major research efforts into default reasoning which do not adopt a logic-based approach to formalizing principles of default reasoning (e.g., Collins & Michalski 1989).

Having selected a neutral term for the phenomenon, let us turn to indicating by example what we are including in the phenomena and what most theories wish to account for. From the statements "Typical linguists speak more than three languages" and "Kim is a linguist", one might draw the conclusion *by default* that "Kim speaks more than three languages." Now, what theorists mean by the phrase "by default" seems at least to imply "not by classical logic from a particular knowledge state that includes just the information mentioned in the premise." The general goal is to allow an agent to make plausible inferences from a knowledge state which might be seen either as (a) accurately describing the real world in which most things are only generally, not always, true, or (b) being only an incomplete description of the real world, or (c) some combination of each. In any of these interpretations, the conclusion does not validly follow in the classical deductive sense, because the premises *may* be true when the conclusion is false. Of course, one *could* say that "concluding by default" that the linguist Kim speaks at least three languages, is justified *because* there is no information which indicates that Kim is not covered by the generalization or that Kim is an abnormal linguist in this regard. However, this presumes a particular *process* for resolving the matter, and we wish not to define the phenomenon in terms that presuppose a particular process solution to it.

Note also that we could have made that first statement about the language skills of linguists in different ways: "Linguists typically/usually speak at least three languages" or "Most linguists speak three languages" or the generic "Linguists speak three languages." We use formulations with "typically" in our examples; different theorists use different formulations, but it is unclear that these all mean the same thing. We return to this topic later, after giving some more background and after putting forward our position. We state our position as forcefully as we can, before we consider some rejoinders that our default reasoning theoreticians might urge upon us.

3. DEFAULT REASONING FORMALISMS

There are a number of formalisms and approaches to default reasoning being pursued in AI. These approaches range from those more geared to implementations (such as default inheritance approaches and Bayesian networks) to those that are presented from a formal or logical stance (such as default logic, circumscription, autoepistemic logic, probabilistic logics, and fuzzy logics).

---

[3] Sometimes a specific term has become associated with a particular theory; thus 'default logic' designates the theory devised by Reiter (1980), and 'non-monotonic logic' has come to be associated with the work of McDermott & Doyle (1980).

Within the latter group it is common to differentiate those approaches which are quantitative (probabilistic and fuzzy logics) from the other qualitative ("symbolic") approaches. The quantitative approaches associate some numerical measure to the "degree of uncertainty" of statements and use this information to account for the default reasoning, while the qualitative approaches make use of the values True/False[4] and have mechanisms for retracting an assignment in the face of new information. It is also common to distinguish those approaches which are syntactic (e.g., default logic) from those which are semantic or model-theoretic (e.g., circumscription). These topics, and rather detailed accounts of each of these systems and their interconnections with the others, are covered superbly by the authors in Gabbay et al. (1993). Although these distinctions are important, we will not be concerned with any of these properties and distinctions in this paper, except incidentally; for our concern is at a different level and its generality is intended to cover all such formalisms. In particular, we want to delve into what, for us, are the ultimate background issues, namely....

4. WHY FORMALIZE DEFAULT REASONING AND HOW WILL FORMALIZATIONS BE ASSESSED?

What are the motivations for formalizing default reasoning and what shall be the measure of success? To answer these questions, we reviewed the statements of researchers who were instrumental in setting the current research strands in the area, scholars who were chosen to write general overviews of the field, and the statements given by authors of introductory AI textbooks. The foundational authors can be seen telling the rest of the community what they thought the justification of the new field of inquiry was, the overview authors are consciously writing to present an overall purpose to the field and were chosen for this task because they are perceived as having the appropriate high-level appreciation of the field, and the authors of AI textbooks are collating the perceived collective wisdom of the field and are teaching the new generation of researchers what is the area of default reasoning. We are thus concerned with what the major writers in the area thought, and continue to think, about these "grander issues."

So we turn now to a presentation of the reasons given by certain theorists and scholars that default reasoning is a necessary component of the AI enterprise, and why it is important to construct formal theories of it as a KR theme. The majority of the quotations can be found in the Appendix of this paper; here we mention only a few representatives to use as a springboard for further discussion.

---

[4] Sometimes also Uncertain, and sometimes other values as well, but if there become too many values then, the symbolic approach blends into the quantitative approaches.

> A key property of intelligence–whether exhibited by man or by machine–is *flexibility*. This flexibility is intimately connected with the defeasible nature of commonsense inference…we are all capable of drawing conclusions, acting on them, and then retracting them if necessary in the face of new evidence. If our computer programs are to act intelligently, they will need to be similarly flexible. A large portion of the work in artificial intelligence on reasoning or deduction involves the development of formal systems that describe this process. [Ginsberg 1987, Intro.; also Ginsberg 1993]

> In everyday life, it seems clear that we, human beings, draw sensible conclusions from what we know and that, in the face of new information, we often have to take back previous conclusions, even when the new information we gathered in no way makes us want to take back our previous assumptions ... It is most probable that intelligent automated systems will have to perform the same kind of (nonmonotonic) inferences. [Kraus et al. 1991]

> Most of what we know about the world, when formalized, will yield an incomplete theory precisely because we cannot know everything – there are gaps in our knowledge. The effect of a default rule is to implicitly fill in some of those gaps by a form of plausible reasoning... Default reasoning may well be the rule, rather than the exception, in reasoning about the world since normally we must act in the presence of incomplete knowledge ... Moreover,...most of what we know about the world has associated exceptions and caveats. [Reiter 1978]

> [The perfect paper on the topic would] build a complex formal model for describing human practical reasoning interactions, show the model is plausible and highlight the role of nonmonotonicity in such a model. The emphasis [would] not [be] on the actual practical modeling of human reasoning (which is a major serious task), but on building enough of a model to distinguish the (what is currently known as) nonmonotonic component from other (human reasoning) components.[Gabbay 1993]

We see in all these quotations (and in all the ones in the Appendix) the view that traditional mathematical or formal logic by itself is inadequate for the task of describing ordinary human activities and human commonsense reasoning; rather, these ordinary human activities rely upon people's ability to employ some form of default reasoning. In consequence, progress on the development of intelligent artifacts (robots and the like) will not advance until more research is done into default reasoning and the ways it interacts with other intelligent activities. It is this human

feature of “non-monotonic flexibility” that needs to be instantiated in these artifacts, in order for them to be correctly viewed as intelligent…or indeed for them even to be useful in carrying out some interesting tasks that involve human-like flexibility. Note especially that *every one* of these quotations invokes the notion of “how humans act and change their beliefs” in their justification of why AI requires a formal theory of default reasoning. We don't mean to say that these researchers are latent cognitive modelers. A more neutral interpretation of this appeal to human abilities might be: "Humans do what they have to do, namely reason about a world in which few statements are universally true. They do it all the time, they do it rather well, and they are at the moment the most comprehensive intelligent agent we have around. So if our artifacts are to play a significant role in a world where hardly anything is always true, this capability of humans is something they must aspire to." Implicitly, it seems, there is a (human) standard that first must be achieved; it is unclear whether there is a notion of "surpassing" this standard for some "better" performance, which is an issue we return to later, when we consider possible responses that some of our default reasoning theoreticians might make to our claims.

But just what *is* this human standard, this “human feature of non-monotonic flexibility"? What are "sensible conclusions" and how will we judge whether or not our artifacts accomplish the task? Surprisingly, there has been very little investigation into this topic, although a number of KR theorists recognize that this is a situation that needs improvement (our emphasis throughout):

> To measure overall progress, one essential question has to be asked: How good are the current NMR [non-monotonic reasoning] theories and systems at formalizing and mechanizing *common sense reasoning patterns* ? [Reinfrank 1989]

> [There is] often [a] profound mismatch between nonmonotonic reasoning in the abstract and the logical systems proposed to formalize it. This is not to say that we should abandon the use of formal NM systems; rather, it argues that we should seek ways to make them model our *intuitive conception of nonmonotonic reasoning* more closely. [Konolige 1989]

> We do not want to depend only on the system developer’s intuitions [in a system that performs default inferences]. *They may be commonly agreed in simple, restricted examples, but in more complicated cases [the system developer’s] intuition is no longer a sufficient guide [to people’s intuitions generally].* [Brewka 1991]

> [W]hat I consider to be the most pressing problem facing the field of nonmonotonic reasoning today [is]: the fact that we have…put ourselves in a position where it is almost impossible for our work to be validated by anyone other than a member of our small subcommunity of AI as a whole…Especially troubling is the recent trend away from formalizing our intuitions, and towards simply getting the 'right*' answers to a small set of simple problems.*…To my mind, the nonmonotonic community needs the ability to *test its hypotheses in a rich environment* far more than it needs more progress on theoretical issues. [Ginsberg 1993]

> Of course, humans have been reasoning nonmonotonically for as long as they have been reasoning at all. …[S]ome more or less disciplined procedures have come into use.…Such [procedures] are not discussed explicitly in this [work], which is devoted to formal constructions for artificial intelligence. But they should not be forgotten, either, as they reveal *a background of human practice with which to compare formal proposals* for eventual computer implementation. [Makinson 1993]

> How are we to evaluate the wide range of approaches in nonmonotonic logic? At the time of this writing, it is possible to compare approaches but not to give enduring evaluations or prescriptions of use. We can ask which of a variety of approaches satisfy a variety of algebraic operations, such as transitivity or contraposition. We can ask whether they support conclusions about both individuals and classes. We can ask how they support revision of knowledge.…Nonmonotonic logic is an area of active research interest. *Researchers are still in the process of gathering examples of human reasoning*, some of which are expressed more conveniently in one [default reasoning] approach than in another.… [Stefik 1995]

There are, at least, some writers on this topic who have heralded a call for grounding the development and validation of default reasoning theories in data in human performance. We agree with these positions and in the next sections develop an even stronger argument based on the connection we see between the justification given for including default reasoning as a goal for AI and the methodology used to validate default reasoning theories. We think that the way to test whether a proposed default formalism is correct is to investigate whether it accurately reflects the results that intelligent human agents obtain. Of course, the mechanism need not obtain the results with precisely the same processes that humans use, but it should capture the important features of the phenomenon. At least, given that it is a theory of human behavior, it should be as accurate for describing this behavior as any theoretical account is in describing the realm it is summarizing and

systematizing. It is this feature that we find lacking in all of the default reasoning literature; and it is this feature that would provide the answer to the issues raised by the just-quoted authors.

5. THE APPEAL TO INTUITION

This brings us to our central position, which we shall investigate further in what follows: We believe that formal accounts of default reasoning are ill-grounded because they do not have a clear delimitation of what phenomena they are to give an account of. In practice what happens is that some researcher has an intuition that such-and-so inference is default-correct and that thus-and-such inference is not default-correct, and then constructs a formal edifice upon these particular inferences. For example, a researcher might say that

Birds typically fly		Birds typically fly
Tweety is a bird		Penguins are birds
Tweety flies	and	Penguins typically do not fly
		Tweety is a penguin
		Tweety does not fly

are correct default inferences, but that

Birds typically fly		Republicans typically are not pacifists
Penguins are birds		Quakers typically are pacifists
Penguins typically do not fly	and	Nixon is a Republican
Tweety is a penguin		Nixon is a Quaker
Tweety flies		Nixon is (not) a pacifist

are not correct default inferences. And an entire formal structure might be constructed on these specific inferences. Yet nowhere did our imagined theorist ever investigate whether people in general agree with these claims about the stated inferences; nor (more importantly) was it ever determined that the *patterns of inference* which these inferences exemplify are viewed by ordinary intelligent agents as having these default-validity properties. And yet it is precisely the validity/invalidity of the *pattern* which is of central concern in determining what is to be included in a formal system of default reasoning. To determine whether the claimed validity of a pattern is in accord with ordinary usage, one must investigate whether all instances of the pattern are judged valid by our intelligent agents.

This brings us to what constitutes "validity" in the default reasoning arena. In the classical deductive case, a specific argument is valid or not depending on whether it instantiates a valid argument *pattern*, which is a matter of whether all instances that have true premises also have a true conclusion. In the default reasoning case we should not look merely at one or a few examples of a

pattern and on this basis decree that a general form is valid. We would wish to know that *all substitutions for the predicates of that example pattern yield a valid specific argument*. Even if a theorist is merely going to consult his or her own intuitions, at least s/he might wish to intuit whether the argument seems valid with nonsense predicates or with symbols in place of the predicates.(Lehmann's (1992) work on defining plausibility logics and then specifying the inferences to be sanctioned by any plausibility logic can be seen speaking to this concern). After all, the default reasoning systems decree that such-and-so *argument type* is valid; and if this is the desired claim then one requires more than the acceptability of one instance of the form. One needs the pattern.

We think that when the position is stated as we have just done, few of the theorists would disagree that *patterns* of inference are important, especially given the sentiments expressed in the quotations given above. But there is often a schizophrenic view about the whole enterprise: researchers cite human performance in default reasoning as their motivation, often their *sole* motivation, but many formal theorists seem either (a) to think that their intuitions about whether a pattern is default-valid/invalid will automatically match everyone else's or else (b) to think that what "ordinary people" believe about the default-validity/invalidity of argument patterns is simply irrelevant to their concern of constructing a formal system. We therefore turn to a consideration of this schizophrenic attitude, and argue against the two just-mentioned reasons for using only one's own intuitions as a foundation for the enterprise. We also believe there are further reasons not to trust intuitions in such fields as default reasoning, and we turn to such reasons afterwards.

5.1 "Everyone (who is rational) Agrees with Me"

One reason our theorists might think it is not necessary to investigate how "ordinary people" do default reasoning is that they believe their own intuitions about how to perform such reasoning are an accurate reflection of how everyone else would do it. That is, since they are human, they themselves automatically manifest the human standard. No special investigation into how people would do the "Tweety/Nixon Inferences" is necessary, because everyone (or everyone who was reasonable) would do it the same way: namely, in the way the theorist intuits it is to be done.

This path is fraught with under appreciated dangers. For one thing, our theorists base their intuitions upon a very small number of examples and yet generalize radically from them (as Ginsberg (1993), for one, notes in earlier his quotation concerning validation). The Tweety/Nixon Inferences are examples of the six or seven motivating examples for the entire enterprise. But note that the conclusion of the second Tweety problem is a fact that we *already know to be true* in our world, independently of the argument: it is a well-known characteristic feature of penguins that they don't fly. So believing this is a valid default *inference* is confounded with *retrieval* of a fact.

And we know from the psychological studies in the deductive logic arena[5] that people are quite prone to judge arguments valid if the conclusion is believed true, regardless of whether the argument is really valid (that is, regardless of whether it exemplifies a valid argument pattern). And similarly, people are quite prone to believe that a specific inference if invalid if they disbelieve the conclusion, again regardless of whether the inference is valid. So at a minimum the theorists should base their intuitions on examples where they do not antecedently know whether the premises and conclusion are true or false.

The focus on intuitions about default reasoning, derived from a small set of examples, and the influence of our own real-world knowledge about the validity of the final conclusion are factors that make it possible to generate what seem to be "counterexamples" for someone's theory or formalism. To take an example, Touretzky, Horty and Thomason, in a paper ominously entitled “A Clash of Intuitions” (1987), discuss the on-path versus off-path preemption problem, using an example about Clyde. Recall that Clyde is both a Royal Elephant and an African elephant, which in turn are each elephants *definitionally*; elephants in turn are *typically* gray things, but Royal Elephants are not gray, *by definition*. The authors ponder Sandewall’s (1986) definition of off-path preemption by engaging in a strategy we advocate: they constructed what was alleged to be a topologically identical network in which the predicates were changed in the following way: George (the substitute for Clyde) is both a Chaplain and a Marine, both of whom are *typically* men; men in turn are *typically* beer drinkers, but Chaplains are *typically* not beer drinkers. What is interesting to us is the authors' speculations on why off-path preemption gives a less intuitively-certain answer in George’s case. First we note that, even in posing what should be an analogous problem, the authors themselves have been subtly influenced by their own real world knowledge: while Royal Elephants were definitionally non-gray, Chaplains were described as (merely) typically non-beer drinkers. Not surprisingly, perhaps, the “intuitions” were no longer clear. Second, the authors observe that "the most relevant missing bit of information is the rate of beer drinking among Marines" versus among Chaplains; and that "one would be better off assuming George is a beer drinker" if the beer drinking was more prevalent among Marines than among Chaplains. What we find interesting about this last speculation is not just the appeal to additional relevant knowledge, but the notion of being "better off" with one assumption over another. Better off in what way? Are there other metrics—perhaps situated, goal-directed ones—that will define which default conclusion is better, "more valid", "more correct" to draw? Touretzky *et al*. appreciate that the problem is much more complex than generally portrayed, and we resonate with this appreciation. We think this makes intuition-based approaches even more difficult to take as the only method. We return to this topic later in the paper.

---

[5] See Evans (1987 and Evans *et al* (1993) for a more general discussion of this topic in various areas of logic.

5.2 Psychologism

The second reason we conjecture that theorists might think it irrelevant what “ordinary people” do with respect to default reasoning is that they believe they are studying (or inventing?) an independently-existing realm, called default reasoning, and that the degree to which "ordinary people" correctly apprehend this realm is simply not relevant. Some people may be good at apprehending this realm but others may be bad, they could say, and so investigation of how people do this apprehension is simply beside the point of discovering the truth about the realm of default reasoning, or to inventing an internally consistent formalism for dealing with notions like "typical" or "generally speaking" and "exception." After all, these theorists might point out, we would not take the word of the “average person” who draws incorrect conclusions in certain syllogisms, who thinks that denying the antecedent in propositional logic is proper, and who cannot see that the probability of a conjunction is never greater than the probability of one of its conjuncts. So why, they might ask, should we accord “ordinary people” any more importance in the realm of default reasoning?

We contend that the two cases are different. The case of investigating how people reason as a way to determine what a nonmonotonic logic *should* admit as valid/invalid arguments is quite different from the case of a similar investigation in classical logic or in probability theory or in arithmetic, and we do not wish readers to believe that we intend our methodological discussion to carry over to them. But we also deny that the methodology of research into the psychology of classical logic and probability theory should carry over to default reasoning.

The view that the *content* of field-of-knowledge-X is whatever resides in the collective psychological states of the population at large is called *psychologism* (with respect to field X).[6] At one time psychologism was a common view regarding arithmetic, geometry, and logic. According to such a view, the content of geometry (to take an example) is what people think, when they are thinking geometrically – at least when suitably generalized so as to account for all people’s thoughts. Logic, according to psychologism, is about those inferences that people judge to be valid, suitably generalized. But Frege’s influential review of Husserl (Frege, 1894; see also Frege, 1884) persuaded many theorists, including Husserl, that this was false of logic and arithmetic. And after this Husserl and Frege independently put forth the view that logic was not a subjective matter but instead was an objective matter of the relations among propositions, predicates, and terms.

Their joint efforts in this matter have made it be difficult to find anyone nowadays who holds psychologism with regards to logic, geometry, arithmetic or probability theory.[7] As a

[6] For a full account of psychologism, and a survey of all the areas to which the term has been applied, see Kusch (1995).

[7] There are some halfway positions such as that advocated by Macnamara (1986) invoking a competence/performance distinction. And Kusch (1995) thinks that psychologism even with respect to logic might be making a comeback.

consequence, when we investigate how people actually reason in deductive logic settings, our conclusions are different than they would be if we believed in psychologism; for according to psychologism, people (as a whole) *cannot* make mistakes. If (almost) everyone reasons in such-and-so way, then by definition—according to psychologism—such-and-so *is* logic. It is only if we reject psychologism in logic that we can say that most people *make mistakes* in logic. Hence, it is discovered that people *make more mistakes* in reasoning with Modus Tollens than with Modus Ponens.[8] (See Evans 1987, Evans *et al* 1993, and Rips, 1994, for more complete summaries of the types of errors made). And in the field of probabilistic reasoning experimenters discover that subjects typically *make errors* in assigning probabilities to conjunctions (Tversky & Kahneman, 1983). But we describe these human behaviors as "making mistakes" or "errors" because we believe that psychologism is *false* with respect to these theories. We instead believe that there is an objective theory for each of these areas which is independent of how people reason about it, and that this objective or normative theory sets the standard with which we can make these judgments about how people make mistakes. (If we did not believe this, then our psychological investigations would not discover people's mistakes in these areas, but rather would be evidence for what was correct in the theory!) Propositional logic, the syllogism, probability theory, and the like, were designed to describe some *independent-of-people purpose*. What is a correct inference in these systems is given by those purposes. (E.g., that whenever all the premises are true, so is the conclusion). In setting the standards and in giving the normative ideal, they thereby make it be possible that there are ways people can go wrong in using these systems, and it then makes sense to investigate when, why, and how people make mistakes.

We believe the case is different for default reasoning. In this case, there is a set of human activities that involve default reasoning. And we are in the situation of trying to construct a theory of rationality which will explain or at least simulate this behavior. Our background belief is that somehow a formal account of default reasoning will play a role in this overall theory. It is precisely this view of the landscape that the theorists cited above and in the Appendix have agreed to. So it seems inappropriate for them now to retract their commitment to "how people act and what default

---

As evidence he cites Harman (1973), Engel (1991), Hurford (1986), Haack (1978), Ellis (1979), Slezak (1989), and Churchland (1987). Much of this new favorable attitude toward psychologism, he claims, is due to Quine (1969).

It might also be noted that in the field of probability theory, there has been a movement toward psychologism also. Especially Gigerenzer and his colleagues have argued that the received (Bayesian) probability theory is flawed and the so-called errors that investigators have found people making in their probabilistic judgments (e.g., those reported by Tversky & Kahneman 1983) are only thought to be errors because of a misplaced attachment to an incorrect view of what probability theory is. See particularly Chapter 5 of Gigerenzer & Murray (1987).

[8] We note, however, that there is a large literature that explains some of these "mistake patterns" as "rational inferences"—predictive of what is likely to be true in the world— given the different types of knowledge that people express in *if-then* form, and some of these knowledge types suppress what classical logic would sanction as valid inferences, and invite what classical logic defines as invalid inferences. This recent "rational analysis" perspective within cognitive brings with it distinctions such as that offered by Anderson (1990) between normative and adapative rationality. Also see the volume by Mankelow & Over (1993).

conclusions they draw." Of course it is a matter for the constructed theory to determine what part, exactly, of the human intelligent behavior is due to the default reasoning portion of an overall theory and what part of intelligent behavior is due to other things. *But the very data that the theory is to cover is determined by the practices of "ordinary people."* It is this feature that our earlier-quoted theorists were referring to when they mentioned testing proposed theories of nonmonotonic reasoning against the "background of human practices" (Makinson), or against the "background of human practical reasoning" (Gabbay).

Of course we do not wish to say that *everything people do* in a default reasoning context must directly be put into a formal theory. We think that there is room in the overall theory to have a place for "mismatches" between theory and phenomena, and in such cases we could say that the phenomenon was outside the scope of the theory or that people made mistakes in default reasoning (as they do in deductive reasoning). This is possible because we view a formal theory as just that: a *theory*. And theories are accepted in part because of their simplicity, their systematicity, and the like. This in turn leaves room for the types of mismatches we mentioned. But before we can construct this default logic theory, we need the data....an empirical description of what patterns of default inference "ordinary people" act in accordance with. (This is a topic to which we shall return later).

6. THE PROBLEM WITH INTUITION

We think our rebuttals are quite strong against the two reasons we conjecture that default reasoning theorists are satisfied to use their intuitions alone in formulating their theories. But additionally, we think there are other reasons not to use intuitions in the specific context of formulating a default reasoning theory, and in this section we consider three of them.

6.1 Ambiguity in Predicates

One reason that it is easy to be misled by intuitions in the particular realm of default reasoning is because the default reasoning examples considered often make use of ambiguous predicates. Some predicates in English are ambiguous between a "capacity" meaning and an "occurrence" meaning. For example, 'fly', especially in the present tense ('Max flies'), is ambiguous between having the *ability* to fly and *actually* flying. The latter sense is usually in the fore when the progressive ('is flying') is used. Presumably it is the former, ability sense which is the concern of the default reasoning examples. But note that even this sense is ambiguous between "ability to self-propelling through the air" and "ability to be transported through the air."[9] In the former sense (normal) robins and (normal) airplanes fly; in the latter sense (all) people and (all) birds and (all) appropriately-sized, appropriately-heavy objects fly. The extended Tweety inferences are difficult

[9] And we can expect that an adequate KR scheme would distinguish between these senses.

to disentangle while holding to a single sense of 'fly': although normal birds fly (first sense), we note that *no* penguin flies (first sense). Of course penguins may be on planes or have rocket-backpacks, and then they'd fly (second sense). But that's no surprise: *all* birds fly (second sense). Indeed *everything* (of medium size and weight) flies (second sense); there are *no* exceptions in this case.

One wants to help the theorists by always reading the premises/conclusion of this example using the first sense: Birds have the ability to self-propel through the air; Penguins are necessarily birds; Penguins do not have the ability to self-propel through the air; Tweety is a penguin; therefore Tweety does not have the ability to self-propel through the air. But this is not where most theorists stop; they instead go on to consider those situations in which certain penguins are exceptions to the claim that they cannot self-propel through the air. Wearing rocket packs, being airline passengers, being tied to flying eagles, and others, have all been proposed as examples of these exceptions. Yet note that none of these in fact are exceptions to the first, preferred, "ability" sense of flies. It makes one wonder, at least concerning those theorists who contemplate exceptions to penguins being exceptions to birds flying, just what exactly do they think *is* the conclusion supposed to mean? One might also note that even when fixing on this preferred, "ability" sense of 'fly', one can still wonder about the ambiguity hidden in the word 'can': does it mean "can as a matter of individual, contingent ability"? Or maybe "can as a matter of evolutionary design"? It would seem that the types of reasoning involved in the two should be different.

Intuitions can get muddled when such ambiguities are not appreciated or acknowledged in the motivating examples. Now, we would not wish to suggest that no examples can be found which will exemplify the desired "default situations" that are called for in presenting and motivating the theory. But we think it behooves the theorists to consider more carefully whether some portion of the intuitions they are using to guess our (or their) intuitions might not be due to the ambiguities in the predicates of the chosen examples.

6.2 What does 'typically' typically mean?

Another place that default reasoning theorists have not adequately addressed whether they have access to the appropriate intuitions is in the interpretation of example default premises. A glance at motivating examples in the literature will reveal that some default premises are worded with the phrase "typically" or "typical", others with "by default", others with "usually"; some use "most", some use "presumably", still others use "normal" or "normally", and some use "generally" or "in general"; and often the mere bare plural formulation is given. Thus we might have any of the following presented when a motivating example is given.

(a) Birds fly.

(b) Typical birds fly.
(c) Birds typically fly.
(d) By default, birds fly.
(e) Most birds fly.
(f) Birds usually fly.
(g) Birds normally fly.
(h) In general, birds fly.
(i) If something is a bird, then presume that it flies.
(j) Birds generally fly.
(k) Normal birds fly.[10]

It is far from clear that these formulations mean the same thing. If, for example, (b) and (k) meant the same, then normal birds would be typical, and conversely; and if (b) and (e) meant the same then most birds would be typical. Yet these do not seem necessarily to be right. *Is* the typical bird the most common? *Are* most birds typical?

In particular the bare plural formulation of the generic in (a) seems quite different from the others. As Krifka *et al*. (1994) have argued, such pure generics are used to express nomic regularities (unlike the other formulations), and such regularities can be nomically true with only very minimal obedience of the instances to the law. For example, the generic sentence *Turtles live to a very old age* is true, despite the fact that the vast majority of turtles are eaten before they are a few weeks old. One explanation of this might be that the sentence means to express that members of the *species* of turtles are *capable* of living long lives, where 'long life' is relative to notions of life spans for other animals. But there are other possible explanations; and in any case the point is merely that such generics are different from the other types of statements. Also, mere universal agreement of instances does not *necessarily* make for a generic statement's being true, for the agreement cannot be an "accidental" matter. Thus, if all 350 existing black rhinos were to have their tails cut off so that each black rhino is a tailless animal, it still would not make *Black rhinos are tailless animals* be a true generic. Our interpretation of generic statements is influenced by what we already know: the inferences that one might draw upon hearing a generic can vary widely depending on one's real-world knowledge. There are a number of further wrinkles with generic statements that make them unsuitable for use as motivating examples in default reasoning examples. For instance, some explanation needs to be given of the fact that *Lions have manes* and *Lions give live birth* are both true, yet no maned lion gives live birth and no lion that gives live birth has a mane.[11] (Only female lions give live birth and only male lions have a mane). Further,

[10] All of the crucial words in these example sentences are highly vague or even ambiguous, but this feature seems even more noticeable in the case of "normal" here.

[11] Two possible explanations were given by a referee of this paper: (a) that *lions have manes* is simply an abbreviation for *some lions have manes*, (b) that *lion* is reserved for male Panthera Leo and *lioness* for a female. In

although *Lions have manes* is true, *Lions are male* is false despite the fact that the class of male lions properly includes the class of the male lions that have manes (e.g., male cubs).

Thus, it is certainly not clear that sentences (a)-(k) above would be represented in the same way in an adequate KR scheme, that is, a scheme which has a place for all (or most) of the concepts we ordinarily use. The question of whether the *same* inferences can plausibly be drawn from each of (a)–(k) itself seems to call for empirical study. It is dangerous for current default reasoning theorists merely to cite these ordinary language arguments and haphazardly invoke these different formulations. For they are then in effect betting that all future detailed, formal KR theories will assign the same meanings to (say) 'typicality' and 'commonalty,' to 'most' and to 'presumably.'

6.3 General Reasons to Distrust Intuitions

Additionally, there is a more general reason why one should be wary of trusting one's intuitions in fields such as this. We find ourselves sometimes surprised by the discoveries in logic and arithmetic, and this shows that we cannot always trust our intuitions about these matters. Furthermore, in related areas experience has shown that intuitions are not always infallible guides. For example, linguists have long warned against trusting the linguistic intuitions of non-native speakers of a language, no matter how apparently fluent they are in the other language. And in the "ordinary language philosophy" of a few generations past, according to which a theorist could consult his intuitions about how terms were ordinarily used and construct a philosophical theory on that basis, it was eventually pointed out that the literature contained contrary claims by the pre-eminent practitioners of the method.[12] After much dispute concerning whether one has any privileged access to linguistic intuitions of this sort, it was decided that we don't. Ordinary language philosophy of this sort faded away and is no longer with us. We should hope that the KR community does not try to reinstate it.

7. THE PROBLEM VERSUS THE PRESCRIPTION

The preceding portion of this paper can be viewed as containing two topics: (1) concerns over how one would find suitable example arguments to test default reasoning theories, and (2) issues concerning what is the justification or grounding for looking at people's actual performance on

---

fact, we do not agree with either of these claims. If one looks at almost any animal book, our general knowledge about lions is captured with such generic statements as *lions have manes* and *lions give live birth*. There is no hint that it is an abbreviation for anything else, or that the words are being used ambiguously. But if one insists, there are equally troublesome problems involving other generics, such as *Cardinals are red* and *Cardinals lay smallish, blue-dappled eggs*. The position that generic statements are really abbreviations or are using the subject terms ambiguously seems to fly in the face of the intent of default logic. For, this seems to be a way of trying to replace the generic statement and its associated exceptions with a statement that has no exceptions.

12 The dispute about ordinary language philosophy that prevailed in the 1940's and 1950's, and its surrounding literature, has been collected in Lyas (1971).

default inferences, or whether indeed there *is* any justification for looking at human default reasoning. Our experience in discussing this paper with others is that, while many people agree with our position on topic (1)—even if they sometimes think it is rather too obvious to point out, our position on topic (2) does not meet with such general approval.

Thus, many theorists are sympathetic with our complaints that researchers do not spend enough time evaluating their proposed sample arguments before they construct a formal edifice that will justify these samples. Many people are supportive of the various recommendations we give that might help to stem the "rush to theory" of certain default theoreticians, such as (a) do not use sample arguments where you already know the truth/falsity of the conclusion on independent grounds, (b) try the argument pattern out with a number of vastly different predicates, (c) try the argument pattern out with nonsense words, (d) try putting a few of the proposed simpler arguments together and evaluate the compound argument, (e) be very wary about using different formulations of the default premise (e.g., 'most' vs. 'typically' vs. 'usually' vs. bare plurals, etc.), (f) be wary about changing meanings of predicates in sample arguments (recall the concerns about 'flies' and 'can').

Indeed, we think that the majority of the default reasoning community would feel favorable towards with these recommendations. Given our earlier quotations from researchers on the topic of validation, some members of the community seem to directly or indirectly agree with the spirit of these recommendations. And probably each member of the community thinks of him- or herself that s/he already follows, at least implicitly, these recommendations in order to get a good grip on what arguments should be considered default-valid.

It is with respect to our views concerning topic (2)—the benefit, if not need, to consider how people really do default reasoning— that some members of the default reasoning community take issue with us. We have claimed in this paper that, unlike classical logic, default reasoning is basically a psychologistic enterprise – that theorists should regard the general goal as to simulate how people reason in a certain type of situation. After all, we have argued, it is human performance in this realm that has defined the phenomenon in the first place. It is this claim that many theorists have objected to, and it is this claim that we wish to further discuss. In the course of this discussion it will become apparent that our position is not so "absolute" as this preliminary characterization suggests.

Let us start with two ways one may wish to deny psychologism in this realm. The first way is to hold to Frege's and Husserl's views of classical logic and mathematics, and simply to deny that default reasoning is in any interesting way different from these. One might say, for instance, that classical logic is the logic of mathematical proofs and consequently is, to some degree, a theory of human activities – a theory of what mathematicians do when they evaluate the soundness of proofs. One might hope or wish that in default logic, as in classical logic, the intuition of the

researcher, combined with the requirement that the formalization be mathematically elegant, will be an adequate guide.

Now, this is certainly a possible attitude to take (although one might quibble over the characterization of classical logic); yet it is not the attitude taken by the field of default reasoning in AI, as can be verified by studying again the quoted statements made by (almost) all of the major researchers in the field. These theorists all say that the goal is to construct artificial systems which will exhibit the level of default-reasoning skill that humans exhibit in the situations that are not governed by hard-and-fast rules.

Of course, the theorists may not really want to be explaining human reasoning behavior, but rather they might view themselves as pursuing some more general style of reasoning. Perhaps, they might say, the human version is just a special case of this more general style, or perhaps the human version is just being used as an analogy.[13] But if this is what some theorists have in mind, there is still a need for some account (independent of a particular system) of what this more general type of reasoning might be like and why it would be a good theory to adopt for artificially intelligent agents. Indeed, it seems quite difficult to think that there is a well-motivated, more general notion of default reasoning than what human default reasoning is.

One argument against the view that "Psychologism is correct with respect to default reasoning" was presented to us by Robert Hadley (personal communication). He suggests a competence/performance distinction so that human plausible inference is to a sound qualitative probability theory as human performance at deduction is to sound formal logic. From this perspective, it is preferable to have artificial agents engage in uncertain reasoning according to some provably correct method (e.g., a qualitative probability theory) than to emulate what may be incorrect patterns observed in humans. This perspective would claim that there *is* an independent-of-people's-performance metric that can be used to evaluate a proposed default system. Indeed, not only is the metric independent-of-people's-judgments on default inferences, but it is also independent of people's ability to get around in the world in general. In this it is rather like classical logic's having the independent-of-people metric of truth-preservingness. Theorists who hold this position are not likely to give the type of quotations we cited earlier and in the Appendix. They are more likely to say such things as:

> ...in order to sharpen our intuitions, it might be best to focus on a few concrete application areas, where we have a clear understanding of when it is appropriate to

[13] But this would be confusing to us, since many of the quotations we examined refer directly to the human skill of default reasoning, rather than analogizing to it or generalizing from it. There are no doubt other authors who do believe in the generalization or analogy. Additionally, it could be argued that even if a system is not meant to model human reasoning, it may need to do human-style default reasoning, since otherwise it might confuse humans with whom it interacts.

> adopt a default rule, or what it means for a formula to be in a set *A* of initial premises, and what it means for a default conclusion to be appropriate....[S]uprisingly little work has been done on this. It would be interesting to consider a number of different interpretations, and consider applications where each interpretation is appropriate. We should not be surprised to discover that different approaches to nonmonotonic reasoning, and different logics, are appropriate for different interpretations of defaults and different applications. [Halpern 1993]

The idea here is that we can have an application arena – for example, a medical diagnosis system – which makes use of default rules. And we can judge whether a particular formalization is better than another by (a) judging whether it allows a system to manipulate, combine, and propagate statistical knowledge in a consistent and well-founded way and (b) calling upon extant knowledge of the application domain to rate the correctness of the conclusion (e.g., the diagnosis). Emulating human behavior may help in getting a system to perform well, but if it comes to a choice between emulation and performing well in diagnosis, we should definitely go for the performance. And as the quotation from Halpern makes clear, using such a paradigm makes it reasonable, even likely, that radically different default rules will be used in different applications.

Similar remarks can be made concerning those theorists who view default reasoning as being some sort of probability theory, and who view classical probability as correct. Here, we have an independent-of-people's performance metric by which to judge the extent to which people are or are not correct: how well they measure up to classical probability theory.[14] This is much the same attitude as taken by Halpern in the quotation above, except for the commitment to classical probability theory.

In these kinds of views of the default landscape, people can "make mistakes" in the pervasive sense whereby most people simply apply the wrong default rules or incorrectly apply the right rules. For, in these kinds of views, people's accuracy in the endeavor is to be judged by the independent-of-people purpose of getting the right diagnosis or in matching Bayesian probability theory.

This attitude toward the field of default reasoning may in fact be what is behind most developers of actual systems that employ default inferences or reason with what we might loosely call statistical information within a specialized domain. This could be the definition that the entire field of AI comes to adopt. Indeed, a current introductory AI textbook (Russell and Norvig, 1995) does not motivate default reasoning as a goal for AI by appealing to human performance in this

---

[14] But of course, this commits such a theorist to believing that default reasoning "really is" probabilistic reasoning, and to believing that probabilistic reasoning "really is" classical. For doubts on the former matter, see Luger and Stubblefield's quote in the Appendix. For doubts on the latter matter, see Gigerenzer & Murray (1987).

arena. Rather, the requirement for default reasoning stems from that fact that "agents almost never have access to the whole truth about their environment" (p. 415). More specifically, the authors' topics do not describe default reasoning as we (and others we have cited) have described it here, but instead focus on "decision-making under uncertainty" with material on probability theory, utility theory, belief networks, and other methods that have proven successful in design of problem-solving systems. Frameworks such as default logic and circumscription are surveyed briefly as "other approaches."

Put another way, default reasoning might cease to be grounded in the notion of "everyday" default conclusions and focus on reasoning from incomplete domain-specific theories and information for expert systems. This said, we find that one still has to admit that most AI *writings* on default reasoning, if not the authors' background beliefs, deny this interpretation and instead presuppose psychologism with respect to default reasoning. Either this trend will continue, in which case our arguments here are relevant, or it will give way to other definitions for which domain-dependent criteria of success will be available.[15]

8. CONCLUSION

What is the proper role in AI theories of default reasoning for studying what people do? It depends, of course, on what those theories aim to do and how they aim to validate themselves. And it was our consideration of these two questions that prompted this position paper. We are not advocating that the formalization of default reasoning cannot proceed without a detailed theory of human default reasoning, let alone all of human cognition. Indeed, as mentioned in the last section, we could instead evaluate how a system performs. What we *are* advocating is that when attempting to construct intelligent artifacts that will work as well as people in ordinary tasks, the formalization should not proceed in a *vacuum* of studies concerning human default reasoning. Our position is that *such data are the very phenomena that default theories are supposed to explain*.

As examples of the sort of benefit that might accrue to default reasoning by paying attention to actual human performance, consider these two discoveries we made in some initial experiments with "ordinary people" (Elio & Pelletier, 1993; Pelletier & Elio, 1993). First, we discovered that although almost of our subjects were happy making this default inference:

a is P, Ps are typically Q, ∴ a is Q,

they were much less willing to make this inference

a is P, Ps are typically Q, Ps are typically R, a is not R ∴ a is Q,

---

15 We wonder, however, whether the need to make "everyday" default conclusions will rise again as a component of language understanding systems.

(even when there was no logical relationship between Q and R). We conjectured that people may be following a heuristic that "when an individual is exceptional in one way, then perhaps it is generally exceptional in other ways as well." In most accounts of default reasoning (Pollock (1987) may be an exception to this), the presence of premises stating that a violates other default rules is irrelevant to the inference at hand. (See Lifschitz 1989, Problem #3). Yet this is *not* the way we found our subjects responding to the simple problems we gave them. In a "rich environment" (cf. Ginsberg, 1993 quotation), the skepticism that one default violation engenders that another default rule might not apply might be sensible. In other work (Elio & Pelletier, 1994), we found that the pattern of "inherited" default inferences depended on whether the concept taxonomy concerned natural kind or artifacts. These sorts of results underscore that one must first determine what the scope of a default reasoning theory is, before deciding whether to include such a strategy within the scope of the formalism. If one wishes the formalism to have wide applicability, as all the theorists we quote advocate, then these *are* types of data which need to be considered when constructing a formal default reasoning theory.

Another finding we discovered for these relatively simple problems was this: Although people are happy to make the inference

a is P, b is P, Ps are typically Q, b is not Q ∴ a is Q,

they are less happy to make the inference

a is P, b is P, Ps are typically Q, b is not Q, a and b are both R ∴ a is Q

(even when R has no logical connection with Q). In our experiments, we varied R – varied the ways the exception object b and the object-in-question a were said to be similar. The more similar that b and a were, the more reluctant people were to conclude that a still followed the default rule. We conjectured a kind of "explanation-based default inferencing" for this finding. Should this kind of consideration—the known or apparent similarity between individuals as it pertains to how they obey default rules—be included in formal accounts of default reasoning? We think certainly so, at least to the extent that it can be given a formal basis. Our point here extends beyond these particular empirical findings to the wide range of empirical data already in the cognitive literature that is relevant to what is called default reasoning in the AI literature. We wish to emphasize the extent to which these kind of findings can be mined for insights into what the scope of the default reasoning is, and the principles that guide humans in making everyday default conclusions.

As we acknowledged earlier, the activity of formalizing default reasoning may be evolving into something other than what the initial motivations suggested. The aim now could be to invent a new kind of logic that will behave in a certain manner independent of the predicates and arguments involved, as deductive logic does. Certain answers will be correct within this logic or formalism; and a failure to give those answers would be an error. If this is the new *raison d'être*, then the output of this logical system could indeed be useful for dictating how an agent should reason in the

absence of extensive domain knowledge. Still, many of the formal accounts employ notions about drawing a default conclusion *because* there is no information to the contrary, or *because* the object in question is not abnormal with respect to the predicate in question. Human data on default reasoning might elucidate how an agent could sensibly go about determining that one or the other of these two somewhat underdefined cases holds, even if the ultimate goal is to construct an independently-existing formal system.

Some of the theorists we quoted earlier remark on the need to validate the extant formalisms. We think this validation should include reference to human performance in plausible reasoning. There is a very large literature both in practical decision theory (see Yates, 1990) and in psychology on probabilistic judgments by people. The seminal work by Tversky and Kahneman (see Kahneman, Slovic, and Tversky, 1982) is well-known and their data on human bias in statistical reasoning are often cited as justification for not grounding formalizations of default reasoning in human performance. However, more updated views of this matter are presented in Tversky and Koehler (1994) and in the edited volume by Nisbett (1993). In the qualitative realm, there are fewer works, but we might mention Hoenkamp (1988), who reports older psychological studies, our own work cited earlier, and related work of Hewson & Vogel (1994). There is also work in the cognitive science realm that grapples with grand-scale theories of plausible reasoning, and not usually from a logic-oriented point of view. Good examples of this is the work done by Collins & Michalski (1989) on plausible inferences and Osherson *et al* (1991) on probabilistic reasoning. Casting the net more widely, work on pragmatic reasoning schemas (e.g., Cheng & Holyoak, 1989) and on the invitation (suppression) of invalid (valid) inferences in everyday language use of the conditional (e.g., Byrne, 1989; Cummins *et al*, 1991) can provide a different and potentially useful perspective even on the mistakes humans make in making deductive inferences.[16] Oaksford and Chater (1994) is good example of the "rational analysis" perspective applied to people's performance on the so-called Wason selection task (Wason, 1968), in which the optimal data to falsify a hypothesis must be identified.

While we have not discussed at length a competence/performance distinction with respect to default reasoning, Pollock (1987) does and argues that reasoning can be compared to using language. Pollock may wish to ground theories of default reasoning in competence, but data on human performance can contribute to determining what human competence is. Current empirical data on reasoning and theoretical accounts of human "error patterns" in probabilistic, inductive, and deductive reasoning may speak to the principles underlying the common-sense, default reasoning that people seem so good at. We consider these sorts of investigations important to the establishment of appropriate methodology in formal theories of default reasoning; but we also

[16] These references to potentially relevant work on default, inductive, statistical, and probabilistic reasoning by people is by no means meant as an exhaustive list.

believe this type of investigation forms a practical research arena. Indeed, much of the formal analyses of, and introspection into, this phenomenon has come primarily *from* the AI community and represents a significant contribution to the other cognitive sciences concerned. The study of the phenomenon of default reasoning can naturally be seen as leading to interconnections with other areas of cognitive science. Almost any activity that one can envision as employing default reasoning can also be seen as invoking other cognitive abilities; and this interplay amongst the various cognitive structures can be used as a way to draw closer ties amongst the various cognitive sciences. There is no need for theoreticians to wait for good empirical and theoretical work on human default reasoning to proceed; it is out there. We hope that issues and arguments we have presented here prompts a closer look at the motivations, definitions, and methodologies that characterize much of the work in formalizing default reasoning within AI, and prompt AI researchers to reconsider the role of human performance in formulating and validating their theories.

ACKNOWLEDGMENTS

We gratefully acknowledge NSERC Research Grants OPG 5525 (FJP) and OPG 0089 (RE). We also thank Aditya Ghose, Robert Kermode, and audiences at University of Edinburgh, the Society for Exact Philosophy, the International Congress for Cognitive Science, and the Canadian IRIS-B5 group for many useful comments. We also thank Robert Hadley for his detailed commentary on a number of our works, and Joe Halpern for a series of wide-ranging discussions on the topic of this paper. Finally, we would like to thank three anonymous referees for their very thoughtful and thought-provoking comments on an earlier draft of this paper.

REFERENCES

Anderson, J. R. 1990. The adaptive character of thought. Lawrence Erlbaum, Hillsdale, NJ.

Besnard, P. 1989. An introduction to default logic. Springer-Verlag, Berlin.

Bibel, W. 1991. "Forward." *In* Nonmonotonic Reasoning: Logical Foundations of Commonsense Reasoning. *Edited* by G. Brewka. Cambridge University Press, Cambridge, MA.

Brewka, G. 1991. Nonmonotonic reasoning: Logical foundations of commonsense reasoning. Cambridge University Press, Cambridge, MA.

Byrne, R. M. J. 1989. Suppressing valid inferences with conditionals. Cognition, **31**:61-83.

Cheng, P. and K. J. Holyoak. 1989. On the natural selection of reasoning theories. Cognition, **33**: 285-313.

Churchland, P. 1987. Epistemology in the Age of Neuroscience. Journal of Philosophy, **84:** 544-553.

Collins, A. and R. Michalski. 1989. The logic of plausible reasoning: A core theory. Cognitive Science, **13**: 1-49.

Cummins, D.D., T. Lubart, O. Alksnis, and R. Rist. 1991. Conditional reasoning and causation. Memory and Cognition, **19**: 274-282.

Doyle, J. 1979. A truth maintenance system. Artificial Intelligence, **12**: 231-272.

Elio, R. and F. J. Pelletier 1993. Human benchmarks on AI's benchmark problems. *In* Proceedings of the Fifteenth Annual Cognitive Science Society Conference. Boulder, pp. 406-411.

Elio, R. and F. J. Pelletier. 1994. On relevance in nonmonotonic reasoning: Some empirical studies. *In* Relevance: American Association for Artificial Intelligence 1994 Fall Symposium Series. *Edited by* R. Greiner and D. Subramanian, pp. 64-67.

Ellis, B. 1979. Rational belief systems. Rowman and Littlefield, Totowa, NJ.

Engel, P. 1991. The norm of truth. Toronto University Press, Toronto.

Evans, J. 1987. Bias in human reasoning: Causes and consequences. Lawrence Erlbaum, Hillsdale, NJ.

Evans, J. St. B. T., S. E. Newstead, and R. M. J. Byrne. 1993. Human reasoning. Lawrence Erlbaum, Hillsdale, NJ.

Frege, G. 1884. The foundations of arithmetic. Translated by J. Austin, 1950. Blackwells, Oxford.

Frege, G. 1894. Review of Husserl, Philosophie der Arithmetic. Zeitschrift für Philosophie und philosophische Kritik 103: 313-332. Extracts translated and reprinted in P. Geach and M. Black, 1952, Translations from the Philosophical Writings of Gottlob Frege. Blackwells, Oxford, pp. 79-85.

Gabbay, D. 1993. Preface. *In* Handbook of Logic in Artificial Intelligence and Logic Programming: Vol. 3: Nonmonotonic Reasoning and Uncertain Reasoning. *Edited by* D. Gabbay, C. Hogger, and J.A. Robinson. Oxford University Press, Oxford, pp. v-xii.

Gabbay, D., C. Hogger, and J. A. Robinson. 1993. Handbook of Logic in Artificial Intelligence and Logic Programming; Vol. 3: Nonmonotonic Reasoning and Uncertain Reasoning. Oxford University Press, Oxford.

Gigerenzer, G. and D. Murray 1987. Cognition as intuitive statistics. Lawrence Erlbaum, Hillsdale, NJ.

Ginsberg, M. 1987. Readings in nonmonotonic reasoning. Morgan Kaufmann, Los Altos, CA.

Ginsberg, M. 1993. AI and nonmonotonic reasoning. *In* Handbook of Logic in Artificial Intelligence and Logic Programming: Vol. 3: Nonmonotonic Reasoning and Uncertain

Reasoning. *Edited by* D. Gabbay, C. Hogger, and J.A. Robinson. Oxford University Press, Oxford, pp. 1-33.

Haack, S. 1978. Philosophy of logics. Cambridge University Press, Cambridge, MA.

Halpern, J. 1993. A Critical Re-examination of Default Logic, Autoepistemic Logic, and Only Knowing, Computational Logic and Proof Theory: *In* Proceedings of the Kurt Gödel Conference, Springer-Verlag Lecture Notes in Computer Science. Springer-Verlag, Berlin, pp. 43-60.

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

Hewson, C. and C. Vogel 1994. Psychological evidence for assumptions of path-based inheritance reasoning. *In* Proceedings of the Sixteenth Conference of the Cognitive Science Society, Atlanta, pp. 403-414.

Hoenkamp, E. 1987. An analysis of psychological experiments on non-monotonic reasoning. *In* Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI87), pp. 115-117.

Hurford, J. 1986. Language and number. Blackwells, Oxford.

Kahneman, D., P. Slovic, and A. Tversky. 1982. Judgment under uncertainty: Heuristics and biases. Cambridge University Press, Cambridge, England.

Konolige, K. 1989. Hierarchic autoepistemic theories for nonmonotonic reasoning: Preliminary report. *In* Nonmonotonic Reasoning. *Edited by* M. Reinfrank, J. de Kleer, and M. Ginsberg. Springer-Verlag, Berlin, pp. 42-59.

Kraus, S., D. Lehmann, and M. Magidor. 1990. Nonmonotonic reasoning, preferential models, and cumulative logics. Artificial Intelligence, **44:** 167-207.

Krifka, M., F. J. Pelletier, G. N. Carlson, A. ter Meulen, G. Chierchia, G. Link 1994 Genericity: an introduction. *In* The generic book. *Edited by* G. N. Carlson and F. J. Pelletier. Chicago University Press, Chicago.

Kusch, M. 1995. Psychologism. Routledge, London.

Lehmann D. 1992. Plausibility Logic. *In* Lecture Notes in Computer Science #626. *Edited by* E. Börger, G. Jäger, H. Kleine Büning, and M. Richter. Springer-Verlag, Berlin, pp. 227-241.

Lifschitz, V. 1985. Closed world data bases and circumscription. Artificial Intelligence, **27**: 229-235.

Lifschitz, V. 1989. Benchmark problems for formal nonmonotonic reasoning, version 2.00. *In* Nonmonotonic Reasoning. *Edited by* M. Reinfrank, J. de Kleer, and M. Ginsberg. Springer-Verlag, Berlin, pp. 202-219.

Luger, G. and W. Stubblefield. 1989. Artificial Intelligence and the design of expert systems. Benjamin/Cummings, Redwood City, CA.

Lyas, C.1971. Philosophy and linguistics. Macmillan, London.

Macnamara, R. 1986. A Border Dispute: The place of logic in psychology. MIT Press, Cambridge, MA.

Makinson, D. 1993. General patterns in nonmonotonic reasoning. *In* Handbook of Logic in Artificial Intelligence and Logic Programming: Vol. 3: Nonmonotonic Reasoning and Uncertain Reasoning. *Edited by* D. Gabbay, C. Hogger, and J.A. Robinson. Oxford University Press, Oxford, pp. 35-110.

Manktelow, K. I. and D. E. Over.1993. Rationality: psychological and philosophical perspectives. Routledge, London.

McCarthy, J. 1977. Epistemological problems of artificial intelligence. *In* Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, pp. 1038-1044.

McCarthy, J. 1986. Applications of circumscription to formalizing common sense knowledge. Artificial Intelligence, **28**: 89-116.

McDermott, D. and J. Doyle. 1980. Non-monotonic logic, I. Artificial Intelligence, **13:** 41-72.

Moore, R. 1985. Semantical Considerations on Nonmonotonic Logic. Artificial Intelligence **25**: 75-94.

Oaksford, M. and N. Chater 1994. A rational analysis of the selection task as optimal data selection. Psychological Review, **101**: 608-631.

Osherson, D., J. Stern, O. Wilkie, M. Stob, and E. Smith. 1991. Default probability. Cognitive Science, **15:** 251-269.

Pelletier, F. J. and R. Elio. 1993. Some Truths about Non-Monotonic Reasoning. University of Alberta Department of Computing Science Tech Report TR93-12.

Pollock, J. L. 1987. Defeasible reasoning. Cognitive Science, **11**: 481-518.

Poole, D. 1988. A logical framework for default reasoning. Artificial Intelligence, **36:** 27-47.

Quine, W. 1969. Epistemology naturalized. *In* Ontological Relativity and other Essays. *By* W. Quine. Columbia University Press, NY.

Reinfrank, M. 1989. Introduction. *In* Nonmonotonic Reasoning. *Edited by* M. Reinfrank, J. de Kleer, and M. Ginsberg. Springer-Verlag, Berlin, pp. vii-xiv.

Reinfrank, M., J. de Kleer, and M. Ginsberg.1988. Nonmonotonic Reasoning. Springer-Verlag, Berlin.

Reiter, R. 1978. On reasoning by default. *In* Proceedings of TINLAP-2, Association for Computatinal Linguistics, University of Illinois, pp. 210-218.

Reiter, R. 1980. A Logic for default reasoning. Artificial Intelligence, **13**: 81-132.

Russell, S. J. and P. Norvig 1995. Artificial Intelligence: A Modern Approach. Prentice Hall, Englewood Cliffs, NJ.

Sandewall, E. 1986. Non-monotonic inference rules for multiple inheritance with exceptions. *In* Proceedings of the IEEE 74, pp. 1345-1353.

Selman, B. and H. Kautz 1989. The complexity of model-preference default theories. *In* Nonmonotonic Reasoning. *Edited by* M. Reinfrank, J. de Kleer, and M. Ginsberg. Springer-Verlag, Berlin, pp. 115-130.

Slezak, P. 1989. How not to Naturalize the Theory of Action. *In* Computers, Brains, and Minds. *Edited by* P. Slezak and W. Albury, Kluwer, Dordrecht, pp. 137-166.

Stefik, M. 1995. Introduction to knowledge systems. Morgan Kaufmann, San Francisco.

Tanimoto, S. 1990. The elements of artificial intelligence. W.H.Freeman, NYC.

Touretzky, D., J. Horty, and R. Thomason. 1987. A clash of intuitions: The current state of nonmonotonic multiple inheritance systems. *In* Proceedings IJCAI-87, pp. 476-482.

Tversky, A. and D. Kahneman. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, **90:** 293-315.

Tversky, A. and D. J. Koehler. 1994. Support theory: A nonextensional representation of subjective probability. Psychological Review, **101**: 547-567.

Wason, P. C. 1968. Reasoning about a rule. Quarterly Journal of Experimental Psychology, **20**: 273-281.

Yates, J. F. 1990. Judgment and decision making. Prentice Hall, Englewood Cliffs, NJ.

APPENDIX

HOW KR RESEARCHERS DESCRIBE DEFAULT REASONING: SOME QUOTATIONS

There is an intuition that not all human reasoning can be translated into deduction in some formal system of mathematical logic, and therefore mathematical logic should be rejected as a formalism for expression what a robot should know about the world. The intuition in itself doesn't carry a convincing idea of what is lacking and how it might be supplied. We can confirm part of the intuition by describing a previously unformalized mode of reasoning called *circumscription,*…We will argue that humans often use circumscription, and robots must too. The second part of the intuition–the rejection of mathematical logic–is not confirmed; the new mode of reasoning …coordinates well with mathematical logical deduction. [McCarthy 1977]

…we routinely make assumptions about the permanence of objects and the typical features or properties of objects, yet we smoothly accommodate corrections to the assumptions and can quickly explain our errors away. In such cases, we discard old conclusions in favor of new evidence. Thus the set of our commonsense beliefs changes non-monotonically. … Our beliefs of

what is current also change non-monotonically. … We must continually update our current set of beliefs. The problem of describing and performing this updating efficiently is sometimes called the *frame problem*. In connection with the frame problem, the conventional view suffers…from monotonicity. …The problem of control is the problem of deciding what to do next. [But there are many problems involved, and] one source of each of these problems is the monotonicity inherent in the conventional view of reasoning. [Doyle 1979]

In artificial intelligence, studies of perception, ambiguity and common sense have led to knowledge representations which explicitly and implicitly embody much information about typical cases, defaults, and methods for handling mistakes.…The possibility of failure means that the formalizations of reasoning in these areas must capture the process of revisions of perceptions, predictions, deductions, and other beliefs.… [P]hilosophers and researchers in artificial intelligence have been forced to face [this issue] because humans and computational models of humans are subject to continuous flow of new information.…[One problem] is the problem of maintaining a set of facts which, although expressed as universally true, have exceptions. … Such…cases include many forms of inferences, default assumptions, and observations.

[McDermott & Doyle 1980]

Research in the theory of commonsense reasoning has revealed a fundamental difference between how universal assertions are used in mathematics on the one hand, and in the area of commonsense knowledge on the other. In mathematics, when a proposition is claimed to be universally true, the assertion includes a complete list of conditions on the objects involved under which the proposition is asserted. But in everyday life we often assert that a certain proposition is true "in general"; we know that there are exceptions, and we can list some of them, but the list of exceptions is not a part of the assertion. …The language of predicate logic has been created primarily for the purpose of formalizing mathematics.…If we want to use that language for representing commonsense knowledge then methods for formalizing assertions about exceptions have to be developed. The study of such methods belongs to the area of *non-monotonic logic*. [Lifschitz 1985]

It has been generally acknowledged in recent years that one important feature of ordinary common-sense reasoning that standard logics fail to capture is its *nonmonotonicity*. … Autoepistemic logic is intended to model the beliefs of an agent reflecting upon his own beliefs. …We are trying to model the beliefs of a rational agent…An autoepistemic logic that meets these conditions can be viewed as a competence model of reflection upon one's own beliefs.…It is a model upon which the behavior of rational agents ought to converge as their time and memory resources increase.

[Moore 1985]

Our long-term goal…is to express these ["common-sense"] facts in a way that would be suitable for inclusion in a general-purpose database of common-sense knowledge. … Common-sense knowledge must be represented in a way that is not specific to a particular application. … Both common-sense physics and common-sense psychology use non-monotonic rules. An object will continue in a straight line if nothing interferes with it. A person will eat when hungry unless something prevents it. [McCarthy 1986]

It is commonly acknowledged that an agent need not, indeed cannot, have absolute justification for all of his beliefs. An agent often assumes, for example, that a certain member of a particular kind has a certain property simply because it is typically true that entities of that kind have that property. .... Such default reasoning allows an agent to come to a decision and act in the face of incomplete information. It provides a way of cutting off the possibly endless amount of reasoning and observation that an agent might perform.... [Selman & Kautz 1989]

All the methods we have examined [Bayesian probability theory, "certainty theory", fuzzy set theory, Dempster/Shafer evidence theory] can be criticized for using numeric approaches to the handling of uncertain reasoning. It is unlikely that humans use any of these techniques for reasoning with uncertainty, and many applications seem to require a more qualitative approach to the problem. For example, numeric approaches do not support adequate explanations of the causes of uncertainty. If we ask human experts why their conclusions are uncertain, they can answer in terms of the qualitative relationships between features of the problem instance. … Similarly, numeric approaches do not address the problem of changing data.… *Nonmonotonic reasoning* addresses these problems directly. … Nonmonotonicity is an important feature of human problem solving and common-sense reasoning. When we drive to work, for example, we …
[Luger & Stubblefield 1989]

It is generally admitted that such [exceptionless] knowledge is not so usual in real world. Rather, people employ rules of thumb, for instance...."Believe that a bird can fly as far as there is no evidence that the bird under consideration is unable to fly" seems to be followed by everybody. Such rules are appealing because they apply frequently while leading to very few mistakes. They appear to be just what is required in everyday life where "things that are generally true" outnumber "absolute truths". [Besnard 1989]

In everyday life, people seem to reason in ways that do not adhere to a monotonic structure. For example, consider the following: "Helen was attending a party...." Here Helen has performed non-

monotonic reasoning. … There is good reason for people to employ such non-monotonic reasoning processes. We often need to jump to conclusions in order to make plans, to survive; and yet we cannot anticipate all of the possible things that could go wrong with our plans or predictions. We must make assumptions about things we don't specifically know. Default attributes are a powerful kind of knowledge, since they permit useful conclusions to be made, even if those conclusions must sometimes be revoked. Here we examine means for AI systems to make defaults of a particular kind [circumscription]. [Tanimoto 1990]

In order to approximate the behavior of a human knowledgeable in some special area of expertise, ...it seems to be necessary to get the system to reason the way human beings do. ...This goal is far easier stated than achieved. What exactly is the way human beings use reasoning, in the first place? Even if there was some uniform way and we had found it, there would still remain the task of casting it into a formalism suitable for computers. The combination of both these tasks leaves us with what seems the only way of approaching the goal: start with some conjecture about the human way of reasoning; cast it into a formalism; test the formalism's behavior in applications, in comparison with human reasoning; if necessary revise the conjecture and start over again; and so forth. [Bibel 1991]

The goal of AI is to improve understanding of intelligent behavior through the use of computational models. One of the few things researchers in this young science commonly agree upon is the importance of knowledge for intelligence. Thus the study of techniques for representing knowledge in computers has become one of the central issues in AI....To formalize human commonsense reasoning something different [from classical logic] is needed. Commonsense reasoning is frequently not monotonic. In many situations we draw conclusions which are given up in the light of further information. [Brewka 1991]

There are lots of nonmonotonic mechanisms. It is not clear how they fit into a coherent thematic view. I believe, however, that it is possible to present a good framework. It must be possible. After all, these logics are supposed to analyze human practical reasoning. We humans are more or less coherent so there is something there! [Gabbay 1993]