Human Benchmarks on AI's Benchmark Problems

Renée Elio
Department of Computing Science
The University of Alberta
Edmonton, Alberta T6G 2H1
ree@cs.ualberta.ca
403-492-5444

Francis Jeffry Pelletier
Department of Philosophy
Department of Computing Science
The University of Alberta
Edmonton, Alberta T6G 2H1
jeffp@cs.ualberta.ca
403-492-0625

Abstract

Default reasoning occurs when the available information does not deductively guarantee the truth of the conclusion; and the conclusion is nonetheless correctly arrived at. The formalisms that have been developed in Artificial Intelligence to capture this mode of reasoning have suffered from a lack of agreement as to which non-monotonic inferences *should* be considered correct; and so Lifschitz 1989 produced a set of "Nonmonotonic Benchmark Problems" which all future formalisms are supposed to honor. The present work investigates the extent to which humans follow the prescriptions set out in these Benchmark Problems.

Keywords: reasoning and problem-solving, decision-making, foundations

I. Introduction

Default reasoning occurs whenever the evidence available to the reasoner does not guarantee the truth of the conclusion being drawn; that is, does not deductively *force* the reasoner to draw the conclusion under consideration. ('Force' in the sense of being required to do it *if* the reasoner is to be logically correct). For example, from the statements 'Linguists typically speak more than three languages' and 'Kim is a linguist', one might draw the conclusion, by default, that 'Kim speaks more than three languages.' What is meant by the phrase 'by default' is that we are justified in making this inference because we have no information which would make us doubt that Kim was covered by the generalization concerning linguists or would make us think that Kim was an abnormal linguist in this regard.

Formally speaking, the term 'non-monotonic reasoning' refers to argumentation in which one uses certain information (the *premises* of the argument) to reach a conclusion, but where it is possible that later adding some further information to those very same premises could make one want to *retract* the original conclusion. The catch-phrase of non-monotonic reasoning is "that new information makes one withdraw previously-made inferences without withdrawing any background premises."

It is easily seen that the informal notion of default reasoning manifests a type of non-monotonic reasoning. More generally speaking, default statements are said to be true about the class of objects they describe, despite the acknowledged possible existence of "exceptional instances" of the class. In the absence of explicit information that any particular object is one of the "exceptional instances," we are enjoined to apply the default statement to the object. However, further information may arrive telling us that this object in fact is one of the "exceptional" ones. This is where non-monotonicity resides in default reasoning.

In artificial intelligence there are two general schools of thought as to how to characterize formally default reasoning. (i) Our background information is associated with a "likelihood" parameter and our new conclusions are modulated accordingly. The most common version of this type is to assign our beliefs or information states a "probability" and to draw conclusions in accord with a probabilistic logic. Another version of this type employs "fuzzy logic." (ii) Our background information is characterized as being "typically true", and we draw conclusions that are treated as 'true', or 'true in the absence of information to the contrary.' The difference between the two versions of default reasoning amounts to whether we explicitly represent our lack of deductive conclusiveness in some *quantitative* way, always attaching some evaluation to each of our beliefs and propagating evaluations to our newly-drawn conclusions. Method (i) enjoins us to do so; whereas method (ii) instead tells us to treat each belief as *qualitatively true* but to be prepared to retract or withdraw conclusions in the face of new information.

In the artificial intelligence literature, drawing conclusions in accordance with method (i) is usually called "uncertain inference" (Shafer & Pearl, 1990), whereas drawing them in accordance with method (ii) is usually called "nonmonotonic reasoning" (Reiter, 1987; McCarthy, 1980; McDermott & Doyle, 1980; Moore, 1985). There have been theoretical studies of uncertain inference in Philosophy (Kyburg, 1988), Computer Science (Bacchus, 1991), Management Science (Yates, 1991) and in Electrical Engineering (Zadeh, 1975). The same cannot be said about the qualitative method (ii), nonmonotonic reasoning. Here the theoretical foundations have been investigated mostly in Artificial Intelligence (see Ginsberg, 1987), but without a consensus on what is the correct underlying logical structure. Indeed, there is even much doubt as to which inferences *ought* to be sanctioned and which *ought* to be disallowed. Partly to ameliorate this problem, Lifschitz (1989) published a list of 25 "Nonmonotonic Benchmark Problems" which gave the answers generally accepted by researchers in the area. All future formal accounts of nonmonotonic reasoning were supposed to be able to yield these answers.

There are many different types of Benchmark Problems in Lifschitz's list, corresponding to the different types of areas in which default reasoning is seen as useful to the AI community. In this paper we are concerned with a subset of these: the "Basic Default Inference" problems and the "Inheritance Inference" problems (see Fig. 1). We retain the original numbering of the problems.

Non-monotonic theoreticians believe that it is *correct* to make default inferences. The background idea was that people use their reasoning abilities "to get along in the world" very well;

if computers could only emulate people in this regard they too would be able to live up to their promise. It was thus proposed that computational areas which actively used formal logical methods would do well to adopt non-monotonic logics as their method (McCarthy, 1986; Kraus, 1990).

1. Blocks A and B are heavy.
Heavy blocks are normally located on this table.
A is not on this table.
B is on this table.

2. Blocks A and B are heavy.
Heavy blocks are normally located on this table.
A is not on this table. B is red.
B is on this table.

3. Blocks A and B are heavy.
Heavy blocks are normally located on this table.
Heavy blocks are normally red.
A is not on this table. B is not red.
B is on this table. A is red.

4. Blocks A and B are heavy.
Heavy blocks are normally located on this table.
A is possibly an exception to this rule.
B is on this table.

11. Animals normally do not fly.
Birds are animals. Birds normally fly
Ostriches are birds
Ostriches normally do not fly
Animals other than birds do not fly.
Birds other than ostriches fly.
Ostriches do not fly.

12. Animals normally do not fly.
Birds are animals. Birds normally fly.
Bats are animals. Bats normally fly
Ostriches are birds.
Ostriches normally do not fly.
Animals other than birds and bats do not fly.
Birds other than ostriches fly.
Ostriches do not fly.

13. Quakers are normally pacifists
Republicans are normally not pacifists
Quakers who are not Republicans are pacifists
Republicans who are not Quakers are not pacifists
« No conclusion to be drawn about Republican Quakers »

14. Quakers are normally pacifists
Republicans are normally hawks
Pacifists are normally politically active
Hawks are normally politically active
Pacifists are not hawks
Non-Republican Quakers are pacifists.
Non-Quaker Republicans are not pacifists.
Quakers, Republicans, pacifists and hawks are politically active «means all combinations of these, including Republican Quakers»

FIG. 1: SUBSET OF BENCHMARKS (FROM LIFSCHITZ'S BENCHMARK PROBLEMS)

The point we wish to emphasize is this: *Despite the acknowledgement by the artificial intelligence community that the goal of developing non-monotonic systems owes its justification to the success that ordinary people have in dealing with default reasoning, there has been no investigation into what sorts of default reasoning ordinary people in fact employ.* Instead, artificial intelligence researchers rely on their introspective abilities to determine whether or not their system ought to embody such-and-so inference. We therefore posed the question: Do people actually reason in the manner prescribed by the non-monotonic logic community?

In this paper, we present results on people's performance on Benchmarks 1-4 and some pilot data on Benchmarks 11-14. This is the first of several investigations we have underway to identify what factors impact plausible conclusions drawn in fairly well-circumscribed problems.

2. Experiment 1: Basic Default Reasoning

Benchmarks 1–4 are called "basic default reasoning" problems. Each of these problems concerns two objects governed by one or more default rules. Additional information is given to indicate that one of the objects (at least) does not follow one of the default rules. We call this the *exception object* (for that default rule). The problem then asks for a conclusion about the remaining object. The Benchmark Answer—that answer accepted by the AI community—is that the existence

of an exception object for a default rule should have no bearing on conclusions drawn about any other object when using that rule. To test the validity of this assumption, we investigated two factors concerning the exception object that intuitively seemed likely to influence plausible conclusions about the object in question: the specificity of information about how the exception object violates the default rule, and the apparent similarity of the exception object to the object in question.

Subjects. Eighty subjects enrolled in an introductory psychology course participated in partial fulfillment of their required experiment participation.

Design. There were two between-subjects independent variables. The first was the *specificity* of the information about the exception object. In Benchmarks 1-4, the manner in which an object violates the default rule is unspecified (e.g., *Block A is not on the table)* We call this the negative form. The positive form of the problem identified a specific state for the exception. The second between-subject variable was *who* the agent solving the problem was supposed to be: a human (actually, the subject) or a robot. Interviews with pilot-study subjects indicated that this made a difference in the kinds of answers generated. We had no *a priori* prediction or intuition about the human vs. robot dimension, but it seemed an interesting meta-cognitive issue to explore.

There was one within-subject variable: *object similarity*. Each subject answered Benchmarks 1-4 presented in a low-similarity version and in a high-similarity version. The low similarity version had sentences corresponding to just those assertions in the original Benchmark. The high similarity version had additional statements describing commonalities shared by the exception object and object in question. Figure 2 illustrates two of the four combinations of specificity and similarity for Benchmark #2.

low similarity / Positive Form /Human

You know	There is a Craftsman electric drill and there is also a Black & Decker electric drill. Electric drills are normally stored in the utility cabinet. The Black and Decker drill is a cordless model.
You also know	The Craftsman drill is on the workbench.

What is reasonable to decide about where the Black and Decker drill is?

high similarity / Negative form/ Robot

Robot knows	Western Construction and ConCo Consulting have each submitted confidential bids for contract work. Confidential bids are normally kept in the Department Head's office. The bid by Conco Consulting was prepared by an outside consultant.
Robot also knows	The bid by Western Construction is not in the Department Head's office. The Western Construction and the Conco bids were considerably lower than the other bids that were received. Both these companies have good track records for consulting work. Their bids were received 2 hours after the deadline date, which was Friday at noon.

What is reasonable to decide about where the Conco bid is?

FIG. 2: ALTERNATIVE FORMS OF BENCHMARK #2.

Two different cover stories were developed for each Benchmark. We counterbalanced which cover story was used as the low similarity version and which was used as the high-similarity version across subjects.

Procedure. Subjects were randomly assigned to receive either human-positive, robot-positive, human-negative, or robot-negative problems. The eight problems (four Benchmarks under two similarity versions) were randomly ordered and presented in booklet form. To lessen the chance that subjects would detect the underlying similarity among the problems, we put one filler problem between each of the randomly-ordered real problems. These filler problems were similar in format and asked for common sense reasoning conclusions. The instructions emphasized that there were

no right or wrong answers to these problems, and that the goal of the experiment was to discover something about how people make (or how robots should make) plausible conclusions in situations for which there is only general information. Subjects generated their own answers and were told that "can't tell" was also an acceptable answer.

Results. We coded subjects' answers about the object-in-question according to one of four answer categories: (a) it followed the benchmark answer, (b) it followed the exception object, (c) it was some other answer, or (d) "can't tell." The benchmark answer for Benchmark 2 (drills) is that the Black and Decker drill is "in the utility cabinet"; the exception-answer is "on the workbench." An answer that would be coded as "other" might be "in the mail between the Dept Head and President" for the contract-bid example.

A repeated-measures ANOVA was performed on the proportion of answers generated in each category for each problem; under this scheme, answer category becomes another factor.[1] There was a significant main effect for the answer category. Most of the time, people applied the deafult rule. The proportion of benchmark, exception, other, and can't-tell answers were 0.585, 0.100, 0.139, and 0.176, respectively. There was also a significant interaction between answer-category and form (positive or negative) and a significant three-way interaction between answer-category, form, and similarity (**F**(3, 228)=3.18, p = .025). Table I gives the data relevant to this three-way interaction.

Table I

	Answer Category			
	Benchmark	Exception	Other	Can't Tell
Positive				
high similarity	.450	.195	.190	.165
low similarity	.605	.090	.090	.215
Negative				
high similarity	.525	.060	.230	.185
low similarity	.760	.050	.045	.140

There are several interesting features about this pattern of data. For low-similarity versions, negative-form subjects gave more benchmark answers than positive-form subjects (0.760 vs. 0.605). It is also striking that positive-form subjects give "can't-tell" as their other answer of choice under the low-similarity forms.

For high-similarity versions of the problems, negative-form subjects tended to put answers into the "other" category (using information from the extra similarity assertions), whereas positive-form subjects tended to say that the object behaved like the exception. Although this may be an artifact of the stimuli for the negative-version (unspecific, negative information about an exception might make it difficult for subjects to say that the object-in-question follows the exception), nonetheless it is a fact that inter-object similarity did increase the proportion of "like the exception" answers given on positive-form problems. Nonmonotonic theories would not predict any effect relating to whether or not we know specifically what is going on with the exception object, since such knowledge tells us nothing about the object-in-question. Further, nonmonotonic theories are not prepared to account for the similarity effect (a a possible exception is the theory of Pollock, 1990.)

We also found a significant problem by answer-category interaction and a significant three-way interaction between benchmark, answer-category, and who the problem-solving agent was: human or robot (**F**(12, 912)=2.27, p = .007). Table II gives the proportion of answers in each answer category as a function of problem-solver and benchmark problem. Looking back at Figure 1, it is clear that the first four Benchmarks are variations on the same theme; these variations are supposed not to alter the application of the default rule. Our results found that answer patterns did differ as a function of Benchmark. In particular, the default rule was applied less frequently in Benchmark 3 than on the other benchmarks, regardless of whether the subject was acting as the problem-solver or specifying what a robot should conclude. Benchmark 3 is interesting because

[1] The model defined by significant main effects and interactions reported here was tested on log-linear transformation of the data, and a chi-square test revealed no significant difference between the predicted and observed data.

each of the objects violates one of the two default rules; our results on this Benchmark suggest that if people find an object atypical in one way, they may find it plausible to conclude it will be atypical in other ways.

The proportions in Table 2 show that Benchmark 3 is different than the other Benchmarks, when subjects specify plausible conclusions for themselves as the problem solvers. And this difference is mostly due to more "other" answers. When subjects give plausible conclusions for a robot as the problem-solver, then Benchmarks 3 and also 2 (where the object-in-question has an additional feature) are different, and these differences are due to more "can't tell" answers. This is what might be called *The Asimov Effect* : people believe that robots should be cautious (and say they can't tell) and not always reason as they would permit themselves to.

Table 2

Problem Solver:	Human Benchmark #				Robot Benchmark #			
	1	2	3	4	1	2	3	4
Answer Category								
benchmark .625	.600	.437	.612	.762	.587	.525	.737	
exception	.175	.137	.175	.100	.075	.025	.025	.037
other	.062	.125	.237	.112	.087	.162	.175	.112
can't tell	.137	.137	.150	.175	.075	.225	.275	.112

3. Pilot Results on Inheritance Reasoning

Benchmarks 11-12 invoke a tree-like hierarchy (ostriches—birds&bats—animals) and statements of typical properties had by members of certain positions in the hierarchy. The question for non-monotonic reasoning concerns which of these properties are "inherited" by the next element up (or down) the hierarchy. The reader is encouraged to draw a network representation of the information given in Benchmark 11 and 12, in which links represent the relevant class-sub-class relationships between nodes, which either do or do not inherit features of their respective parents. This will make the preliminary results we present here easier to understand.

The procedure was similar to the study described earlier, where different cover stories were generated for these problems, but we will use the words in the benchmarks to summarize our findings. Nearly 90% of the subjects concluded that animals-other-than-birds could not fly in both Benchmarks 11 and 12, and about this many concluded that birds-other-than-ostriches could fly in benchmark 11. This agrees with the Benchmarks. However, only 53% of subjects concluded that birds-other-than-ostriches could fly in Benchmark 12. Note that the only difference between Benchmarks 11 and 12 is the additional mention in #12 of the subclass "bats" and the fact that they normally fly. Most of the remaining subjects concluded that birds-other-than-ostriches *cannot* fly. What is unusual about this finding is that the mere existance of an extra hierarchy node (bats) influences a decision about a node below it in the hierarchy and on a different inheritance path altogether. We have no account for this result at the present, other than that there may be some influence of how subclasses violate their parents' rule (i.e., since bats violate the animal rule, then maybe birds-other-than-ostriches violate the bird rule), and plan to investigate this result further.

Benchmarks 13 and 14 concern hierarchies in which there are "conflicts"; that is, there is more than one way to traverse the hierarchy, and doing it one way leads to a conclusion opposite to the one generated in traversing it the other way. Again the reader is encouraged to draw the hierachy relating these subclasses. Our preliminary data indicate that subjects behave in accordance with most nonmonotonic theories, which enjoin us to to follow the defaults of the most specific groups to which one belongs. Quakers who are not Republican are pacifists; Republicans who are not Quakers are hawks. Furthermore, if an obect belongs to two distinct groups with differing but overlapping defaults, then that object should at least possess all the defaults that the distinct groups have in common. This means that, for Benchmark 14, our subjects concluded that Republican

Quakers are politically active. Many default mechanisms find it difficult to obey this desideratum, unless it is known how many paths there are and that they will intersect. The other interesting finding concerns the conflict node in these hierarchies: the benchmark answer is that no conclusion can be drawn about whether Republican Quakers are hawks or pacifists. Half of our subjects do generate an answer. For these subjects, half generate the answer that they are hawks, and half generate the answer that they are pacifists. In general, it is interesting that for this sort of non-monotonic reasoning problem, "can't tell" is not a preferred answer, despite the fact this is the preferred answer of every nonmonotonic theory.

4. Conclusions

Non-monontonic logics define what are plausible conclusions in these simple situations, drawing their justification for this from people handle defaults and exceptions. and draw their justification from how it seems that people handle defaults and exceptions. From these data, it seems that people's plausible conclusions about defaults and exceptions are influenced by differences in the amount of information available about the objects (i.e., differences among the scenario specifications of Benchmarks 1-4), are influenced by the specificity of information about the exception, and are influenced by the apparent similarity between objects that might be governed by the same rules. The more that is known about an exception, the more plausible it may seem that another object behaves like it. To us, this suggests an aspect of plausible reasoning that is missing from current non-monontonic theories, namely what kinds of information are relevant to applying default rules. The issue of what is relevant knowledge is only now being examined in the non-monontonic community. One interpretation of our findings is that people do not reason about defaults and exceptions as formal rules to be manipulated: they will put themselves in "problem-solving mode" and integrate *all* the information presented in some way to generate a plausible conclusion. This suggests that it may be difficult to develop robust models of non-monontonic reasoning without some goal-directed component, that in turn determines what kind of information is relevant to the application of a default rule.

It is unclear whether subjects in our study cared that the rules included the term "normally" or whether they would have behaved any differently if that term were omitted. We have debriefing data on what subjects think terms like "normally" or "typically" mean, but have not analyzed that data yet. Futher, there are other factors known to influence deductive reasoning (e.g., premise order effects, belief biases; see Rips, 1990) that are yet to be examined in this domain. We hope the further empirical work in this area can lend some plausible guidance to formalizing context effects in non-monontonic theories.

5. References

Bacchus, F. (1991). Representing and reasoning with probabilistic knowledge. Cambridge: The MIT Press.

Ginsberg, M.L. (1987). *Readings in Nonmonotonic Reasoning* (Los Altos, CA: Morgan Kaufmann).

Kraus, S., D.Lehmann, M.Magidor (1990) "Nonmonotonic Reasoning, Preferential Models, and Cumulative Logics" *Artificial Intelligence* **44**: 167-207.

Kyberg, H. (1988). Probabilistic inference and non-monotonic inference. *Proceedings of the Fourth AAAI Workshop on Uncertainty in AI*. pp. 229-236.

Lifschitz, V. (1989). "25 Benchmark Problems in Nonmonotonic Reasoning, v. 2.0" In M. Reinfrank, J. de Kleer, & M. Ginsberg (Eds.) Nonmonotonic Reasoning. Berlin: Springer). pp 202-219.

McCarthy, J. (1980). "Circumscription -- a Form of Non-Monotonic Logic" *Artificial Intelligence* **13**: 27-39.

McCarthy, J. (1986) "Applications of Circumscription to Formalizing Common Sense Knowledge" *Artificial Intelligence* **28**: 89-116.

McDermott, D. & J.Doyle (1980) "Non-monotonic Logic, I" *Artificial Intelligence* **13**: 41-72.

Pollock, J. (1990). "A Theory of Defeasible Reasoning" *International Journal of Intelligent Systems*.

Reiter, R. (1987). "Nonmonotonic Reasoning" *Annual Reviews of Computer Science* **2**: 147-187.

Shafer, G. & J. Pearl (1990) *Uncertain Reasoning* (Hillsdale,NJ: Lawrence Erlbaum).

Rips, L. J. (1990). Reasoning. *Annual Review of Psychology. 41*, 321-353.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems, 1,* 3-28.