HUMAN PERFORMANCE IN DEFAULT REASONING

Francis Jeffry Pelletier
Department of Philosophy
University of Alberta
Edmonton, Alberta
Canada T6G 2E5
jeffp@cs.ualberta.ca

Renée Elio
Department of Computing Science
University of Alberta
Edmonton, Alberta
Canada T6G 2H1
ree@cs.ualberta.ca

I. Background

There has long been a history of studies investigating how people ("ordinary people") perform on tasks that involve deductive reasoning. The upshot of these studies is that people characteristically perform some deductive tasks well but others badly. For instance, studies show that people will typically perform MP ("modus ponens": from 'If A then B' and 'A', infer 'B') and bi-conditional MP (from: 'A if and only if B' and 'A', infer 'B') correctly when invited to make the inference and additionally can discover of their own accord when such inferences are appropriate. On the other hand, the same studies show that people typically perform MT ("modus tollens": from 'If A then B' and 'not-B', infer 'not-A') and biconditional MT badly. They not only do not recognize when it is appropriate to draw such inferences, but also they will balk at doing them even when they are told that they can make it. Related to these shortcomings seems to be the inability of people to understand that contrapositives are equivalent (that 'If A then B' is equivalent to 'If not-B then not-A'). [Studies of people's deductive inference-drawing abilities have a long history, involving many studies in the early 20th century concerning Aristotelian syllogisms. But the current spate of studies draws much of its impetus from Wason (Wason, 1968; see also Wason & Johnson-Laird, 1972).] The general conclusion seems to be that there are specific areas where "ordinary people" do not perform very logically. This conclusion will not come as a surprise to teachers of elementary logic, who have long thought that the majority of "ordinary people" are inherently illogical and need deep and forceful schooling in order to overcome this flaw.

There has been a similar history–although somewhat shorter–of studies investigating how people perform on tasks that involve probabilistic reasoning. The upshot of these studies is that people characteristically make a number of errors because they do not pay attention to certain relevant information and because they give undue importance to certain kinds of irrelevant information. The study of the types of errors people make on this probabilistic reasoning has come to be called "heuristics and biases" because the underlying assumption is that people have inbuilt or inherent biases as part of their mental equipment that causes them to ignore or give undue status to some information. Usually it is claimed that these biases stem from people having certain rule-of-thumb heuristics which allow them to evaluate information quickly and efficiently in "normal" circumstances but which will lead people astray in other circumstances. (Much of this work stems from Kahneman & Tversky: see Kahneman *et al* 1982 and Evans 1987).

Using a similar methodology, researchers have found that people's decision-making reasoning deviates from the prescriptions of decision theory in a number of regards. Many of the conclusions reached here mirror the "heuristics and biases" approach used in the probabilistic reasoning studies: people simply have tendencies to ignore certain information because of the (evolutionary) necessity to make these types of decisions quickly. This gives rise to "biases" in judgments concerning what they "really" want to do. (See Baron 1994 and Yates 1990).

So, there is a long tradition of finding "ordinary people" wanting in their abilities to perform logical inferences, whether in deductive logic, probabilistic reasoning, or decision making. Yet perhaps we can see something of a difference in the three cases. In the second and third cases there is no universal agreement as to what the "correct" underlying theory is. In the case of decision making there are alternatives to decision-theoretic formalizations. And in the case of probabilistic reasoning there are at least the two alternatives of subjective probability and frequency theories. Some researchers have argued that people are *not* in fact so bad at probabilistic reasoning when the information is presented in the right way and we recognize what the "really right" conclusions are. Gigerenzer and his colleagues (see Gigerenzer 1991, 1994, 1998 for example) think that all the flaws discovered by researchers in this tradition are due to their adherence to a false theory of probability–subjective probabilities rather than the correct frequency theory. And Cohen (1981) proposed a new system of probability theory that was intended to capture the sort of probabilistic reason that people actually follow. (His "Baconian probability theory").

In the probability theory and decision theory cases there is disagreement over what the normatively correct underlying theory is. Hence, when people's performance is measured there can be questions about whether they are performing "correctly" or not. If Theory$_1$ is the normatively right theory for the field, then they are performing badly; but if Theory $_2$ is the normatively right theory in the field, then they are performing correctly. Related to this is yet another idea: One might say that in certain fields there simply *is no other standard* than what people in fact do. In determining how North Americans distinguish green from blue, for example, there is no real concept of right and wrong in the task. Some people draw the boundary in one place, some in another place; we can tell whether there are identifiable subgroups who all draw it at one location while other subgroups draw it in another place; and we can tell whether someone is different from the majority in where s/he draws the line. But there is no real notion of "draws the blue/green boundary incorrectly"; all that exists is how people in fact do it. This attitude is often called "psychologistic" because it locates the object of the study (or the normativity of the theory) in the psychology of people and denies that there is any "external standard of correctness" for the field.

One can imagine that decision theory and even probability theory are psychologistic in this way: there is no notion of "correct inference" other than what people actually do. The correct probability theory (for example) could be some sort of generalization of how people actually reason probabilistically; and therefore when we say that someone is making a mistake in probabilistic reasoning, all we mean is that the person deviates from what people normally do in this regard. Cohen's (1981) view is essentially of this nature; he denies that people as a species can be shown to be irrational because 'rationality' is the sort of notion that is *defined* as what people's underlying capacities have them do in the relevant circumstances.

It is somewhat more difficult to imagine the psychologistic response being made in the realm of deductive reasoning. The reason for this is that we in fact have a perfectly clear "outside standard" of what counts as correct deductive reasoning, namely, *preserves truth from premises to conclusion.* If there were agreement in the probability or decision theories about what the "outside standard" is, then the psychologistic response would be seen as less plausible in those fields also. But once the possibility of a psychologistic attitude is brought up in the deductive arena, we will find some theorists arguing for a "relevance logic" or a "fuzzy logic" or … , on the grounds that "this is the way people *actually* reason".

II. Introduction

Non-Monotonic Logics is the name given to any logic (i.e., any formal, symbolic system) with this feature:

$S_1, S_2, \ldots S_n \vdash C$ and yet it is false that $S_1, S_2, \ldots S_{n}, S_{n+1} \vdash C$

That is, it is possible to go from a valid argument to an invalid argument merely by adding more information to the premises (and not deleting any other premise). This formal property is related to properties of counterfactual conditionals (where adding more information to the antecedent of such a conditional can make the truth-value of the entire conditional change from true to false), and to *prima facie* reasoning in ethics and law (where conclusions of such reasoning tell us what is right or legal in the absence of contravening factors). And many researchers have claimed that similar patterns of reasoning occur in many different fields. As a whole, this type of reasoning (whether or not it is formalized so as to exhibit the non-monotonic logic feature) is called "default reasoning".

Probability theory allows for the probabilistic value of "A, given B" to be greater than the the value of "A, given both B and C", and that to be greater than the value of "A, given B and C and D", etc., so there are many attempts to accommodate all default reasoning within probability theory. (Pearl 1988, Bacchus 1991). Because this way of looking at default reasoning presumes to assign a numerical value between 0 and 1 to sentences (given a background), this is called a *quantitative* theory of default reasoning. Besides probability theory, another quantitative theory of default reasoning is fuzzy logic.

Opposed to the quantitative theories are the *qualitative* theories. These theories do not assign some number between 0 and 1 to sentences, but rather give sentences one of two values: either True or False. (Or sometimes a third value, Uncertain/Undefined, is allowed; and sometimes a fourth value. [See Belnap 1978, where the four values are: True, False, Both True and False, Neither True nor False]. But if there are too many possible values then such a qualitative theory slides into a quantitative theory.) Classical examples of the qualitative theories of default reasoning are Reiter's (1980) *default logic*, McCarthy's (1980) *circumscription*, and Moore's (1985) *autoepistemic logic* (among others).

Even within just the qualitative or quantitative camp, there are a variety of differences among the different theories, both in their overall outlook and methodology and in the class of inferences that they each sanction as being valid default arguments. It is to this issue that we wish to address our next remarks.

III. How Should We Decide on Valid Default Inferences?

In humans, not all of our information about the world is stored in an explicit form in our minds; instead much of it is *implicit* in that it is a consequence of, or follows from, other information which *is* stored explicitly. The reason for this psychological truism is that it would be incredibly inefficient, incredibly wasteful of memory, etc., if everything that we knew were to be mentally entered in "longhand" for our inspection. Instead only some of our knowledge is explicitly stored, and much other information can be recovered or inferred from this explicit store. For instance, none of us has explicitly stored the fact that there is no roadrunner in our refrigerator; yet we know this nonetheless. And the way we know it is that it (somehow) follows from explicitly stored facts about our refrigerator and the locations of roadrunners. The goal of the constructors of "knowledge bases" is to mimic this manner of storage. It is commonly believed by these researchers that this is the only way that major advances will occur in computational information storage.

One of the ways we infer these implicit pieces of knowledge is by default reasoning. Most researchers in this area think this is the most pervasive form of reasoning in our everyday life. It occurs whenever we wish to draw new conclusions from old information, where the old information does not *deductively* guarantee the truth of the new conclusions. For example, from our general knowledge that birds typically can fly, we conclude that our neighbor's new pet bird can fly.…despite the acknowledged possibility that he might have managed to buy a roadrunner. Two quotations taken almost at random from researchers in the area show the attitude which is pretty universally adopted.

> 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)
>
> 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)

We see here that the artificial intelligence community (or at least the knowledge representation subdivision of it) believes that humans explicitly store only a portion of the information that they know, and that when called upon, they typically use some sort of default reasoning to infer the rest.

What might also be noted in these quotes, and any other quotes that one might wish to bring up from this literature, is that they all justify the search for a formal theory of default reasoning by *appeal to human performance*. "It is because humans do such-and-such task so well.…" or "People can get around in the world with no problems…" are typical of the justification given for the whole enterprise. This is *much* different than the case of logic or arithmetic where there is some "external standard" that is independent of how humans reason logically or arithmetically. In the cases of logic and arithmetic we would probably *not* wish our artificial agents to mirror human performance, because we have the desire that these agents should compute the answers *correctly according to the external standard.* But in the case of default reasoning there is *no* standard which is external to how humans perform their default reasoning. Another way of putting the point, using the language developed above, is that default reasoning (unlike logic and arithmetic) is *psychologistic* in nature; it is by definition the study of how people perform on certain types of problems and in certain types of situations. *There simply is no other standard against which to judge how good some abstract system of default inference is, nor any other standard with which to compare artificial agents that embody such a system.* And so it follows that any judgment of the correctness of a proposed system of default reasoning will be as a test of the system against the way humans perform on that reasoning problem. (This outlook on default reasoning, augmented by many quotations from AI researchers who would appear to believe the same, can be found in Pelletier & Elio 1997).

IV. The Non-Monotonic Benchmark Problems

Despite the fact that all the researchers in the field appeal to the goodness of human performance as their justification for employing default reasoning mechanisms in artificial agents, the truth is that none of them in fact have ever investigated how people actually employ default reasoning. A consequence of this is that the different proposed formalisms validate different sets of inferences, and there is no agreed-upon method to decide which inferences should be sanctioned as "really legitimate." Recognizing that there was no accepted background for finding the extent of legitimate default inferences, Lifschitz (1989) set out a number of inferences that were supposed to be valid in any proposed default reasoning system. Different groups of these problems were addressed to different aspects of the default reasoning process, so that perhaps not every reasoning system needed to accommodate all problems; but for any area that a system proposed a mechanism, it was to be able to deal at least with the inferences relevant to that area. We call these problems "the Benchmark Problems", and the accepted answers that were proposed for these problems in Lifschitz (1989) the "AI answers" to the Benchmark Problems.

Although one might accept the legitimacy of establishing a set of benchmark problems in this way because these are the areas in which default reasoning is to be employed by artificial agents, one might nonetheless wonder about the legitimacy of determining the AI answers as a matter of agreement among the various researchers. After all, if it is true that the realm of default reasoning is psychologistic and that therefore the correct answers are determined by the way "ordinary people" will (on the whole) use the method, then the fact that some elite subgroup of people think the answers should be such-and-so is not a good justification. For one thing, their opinions might be colored by how their systems perform. More importantly, an individual's intuitions are not a reliable guide to how the population as a whole treats the phenomenon. In Pelletier & Elio (1997) we have investigated the various reasons researchers give for allowing their own intuitions to be their sole guide in this regard and for not engaging in large-scale investigations of how it is manifested in the population as a whole. We did not find any of these reasons very compelling, and recommended that researchers undertake such studies in order to determine the correct direction for their formal theories to follow.

The present paper sets out the preliminaries for such work, and gives some very tentative results. Despite the tentative nature of the results, we think they should give default reasoning researchers pause in their confidence that they have in fact fathomed the true nature of the reasoning process they are trying to model.

We will discuss only one area here from Lifschitz's list of Benchmark Problems, namely the ones he calls "Basic Default Inference", his Problems #1–#4. As stated in his article they are:

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.

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.

A is red.

FIG. 1: "BASIC DEFAULT INFERENCE" (FROM LIFSCHITZ'S BENCHMARK PROBLEMS)

Each of Problems 1-4, what Lifschitz called the basic default reasoning 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 refer to this as the *exception object* (for that default rule). The problem then asks for a conclusion about the remaining object. We refer to this as the object-in-question.

It is clear from Figure 1 that the four default reasoning problems are variations on the same theme. Problem 2 includes the extra information that Block B is red, but we are still to conclude that it is on the table. Problem 3 mentions two default rules: the exception object violates the default concerning location, and the object-in-question, Block B, violates the default rule about color, but we are still to conclude that Block B follows the default location rule. Problem 4 states only that Block A *might* violate the default rule, in contrast with Problem 1, which states with certainty that there is an exception to the rule.

For all these problems, the given default conclusion is that Block B obeys the default rule concerning location. According to the AI answers – that is, according to the collective wisdom of researchers into nonmonotonic theories – the existence of an exception object for a default rule, or additional information about that exception object, should have no bearing on a conclusion drawn about any other object when using that rule. Extra information about the object in question itself (e.g., Block B's color) should also have no bearing on whether a default rule about location applies. And being an exception object for some *other* default rule should have no bearing on whether it does or does not follow the present default rule.

An implicit but important assumption here is that a logic for manipulating assertions of this form can be developed for non-monontonic reasoning without regard for the semantic content of the lexical items. Given these formal problems about blocks and tables, it seems easy to accept the idea that object color should have no "logical" bearing on whether a default about object location is applied. Yet it is equally easy to imagine real-world scenarios in which it makes (common) sense that an object's color might be predictive of or at least related to an object's default location (e.g., how an artist or designer might organize work items in a studio). We do not wish to confound people's *logical* abilities with their ability to "look up" information they have stored in memory. And so we would want to test them on scenarios that they have some "feel" for but which they have no stored information about.

V. An Experiment into Basic Default Reasoning

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. It is well known (Evans 1987) that posing information in negative form is more difficult for people to deal with in such tasks as drawing inferences than is the same information posed in positive form. Thus we wished to investigate whether the presence of an explicit negation (as the Benchmark Problems had) caused more difficulties in processing these problems than giving positive information concerning the object (which would entail the negative information). The negative form can be seen to "unspecific" with respect to information conveyed: saying that a block is *not* on the table does not say where it is, but saying that the block is on the floor not only implies that it is not on the table but also gives the specific information as to its location. Interestingly, the issue of how negation should be handled in the Prolog programming language is also an issue of some complexity, perhaps mirroring the difficulties that people have in using negative information.

Most people who have thought about default reasoning have considered the possibility that the willingness of someone to draw a given inference perhaps is dependent on the specific lexical items or situations being discussed, and it therefore has nothing to do with logic – that is, nothing to do with the logical form of the inferences to be drawn. In particular for Problems 1-4, one might speculate that the amount of perceived similarity between the exception object and the object-in-question would influence whether subjects thought the object in question should behave like the exception object or should follow the default rule. Thus one of the variables to be manipulated was how much similarity was stated to hold between the two objects.

Subjects. Eighty subjects enrolled in an introductory psychology course participated for partial credit towards an optional experiment-participation component of their course.

Design. There were two between-subject 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 non-specific form. The specific form of the problem identified a particular state for the exception (e.g., *Block A is in the cabinet*). 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. (That is, people might take a different view about the inferences *they* would make with these default logic questions were *they* in the scenario described in the experiment from those inferences they would want or expect *an intelligent robot* to produce). 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 both a low similarity and a high-similarity version of Benchmarks 1-4.The low similarity version had sentences corresponding to just those assertions in the original Benchmark. The high similarity version had additional statements describing commonalties shared by the exception object and object in question. Figure 3 illustrates two of the four combinations of specificity and similarity for Benchmark #2.

Low similarity / Non-specific 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 not in the utility cabinet.

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

High similarity / Non-specific 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 in the auditor'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.

Question: What is reasonable to decide about where the ConCo bid is?

Figure 2: Alternative Forms of Benchmark 2

It is important to appreciate that the additional statements in the high-similarity condition do not enjoin us to conclude that the two objects behave identically with respect to the default rule. On the other hand, we designed these assertions under the guiding principle that they could, conceivably, constitute an explanation as to why an object might not obey a default rule. For example, in the "contract bids" cover story, the extra assertions describe a set of circumstances that could explain why a particular bid is not where it is supposed to be by default: it came in late, it was a very low bid, and it comes from a good company. Maybe these facts together constitute a reason why it is sitting in the auditor's office. They certainly don't *entail* the conclusion that any bid should not follow the default rule and a problem solver would have to do a good bit of work to weave these into an explanation.

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-specific, robot-specific, human-nonspecific, or robot-nonspecific 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 benchmark problems. These filler problems were similar in format and also 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 AI answer, (b) it followed the exception object, (c) it was some other answer, or (d) “can’t tell.” The AI 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. We converted the subject data into the proportion of answers generated in each category for each problem; under this scheme, answer category becomes another factor.

Let us take a first quick look at the percentage of answers that followed the AI answer:

Problem	%AI Answer
1	78%
2	75%
3	56%
4	77%

Table I: Percent of answers agreeing with AI Answer on the Benchmark Problems

It is clear that Problem #3 is significantly different from Problems #1, #2, and #4. (All results we describe as distinct are significant at the p=.05 level). This suggests that the more default rules an object is known to violate, the more “generally exceptionable” the object is, and the more likely the object is to violate other defaults.

Turning our attention to the issue of similarity we see:

Condition	%AI Answer
Hi Similarity:	64%
Lo Similarity	80%

Table II: Percent agreeing with AI answer as a function of Similarity

Although this effect is qualified by an interaction with NEG, this suggests that the object-in-question is more likely to "follow" some other, exceptional object if additional information creates a scenario under which the objects are *seen* as similar – even if this information does *not* state, imply, or "make it really more likely" that the object-in-question should also be an exception.

As remarked, there is an interaction between the degree of perceived similarity of the exception object and the object-in-question, on the one hand, and whether the information given is presented in a negative form ("…the object is *not* …") or in a positive form ("the object *is*…") that implies the negative form.

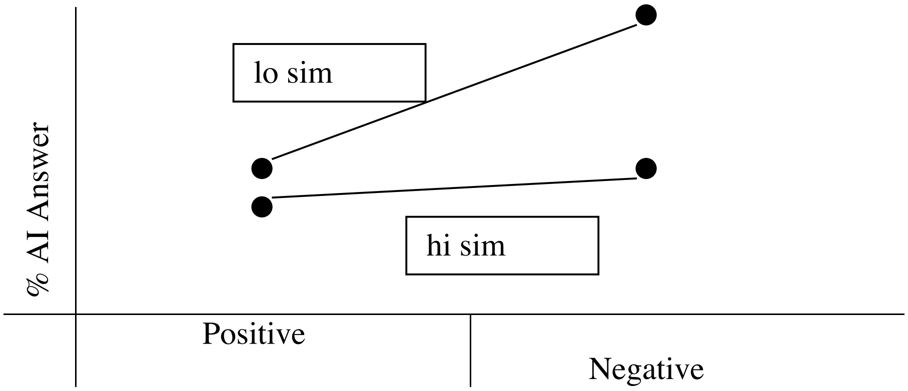

Figure 3: %AI answer's interaction between Similarity and Pos/Neg

Not only is the Low Similarity condition generally more likely to yield the AI answer, but also Positive and Negative affect the Hi/Low Similarity differently. Presumably this has something to do with the fact that negation amounts to some sort of "lack of specificity" about the object (we know that it is 'not manifesting some condition' but we do not know which condition it *is* manifesting), and when this is added to the Low Similarity condition (we do not know anything about whether the object-in-question shares any properties with the exception object) we will more often ignore the exception object's possible influence on the object-in-question.

There was no main effect for either POS/NEG or for Robot/Human, although there was a significant interaction between the two. Subjects more often gave the AI answer for Humans when the information was POS than when it was NEG, but they more often gave the AI for Robots when the information was NEG than when it was POS. Figure 4 displays this result:

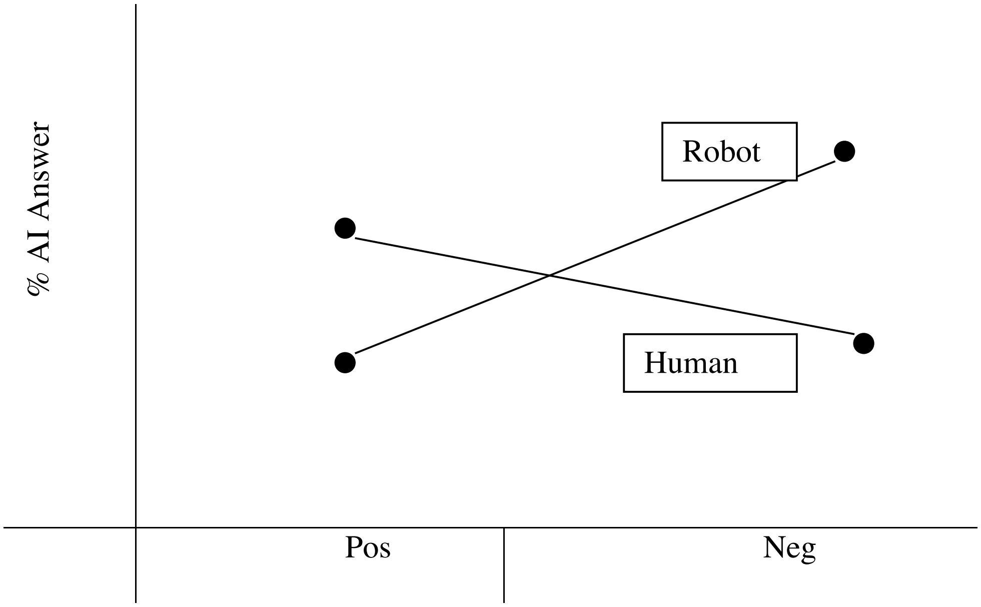

Figure 4: Interaction of POS/NEG with Human/Robot on % of AI answers

We will return to this very puzzling result shortly. But first we will take a somewhat different look at the data collected in this experiment.

Recall that we coded subjects' answers about the object-in-question according to one of four answer categories: (a) it followed the AI answer, (b) it followed the exception object, (c) it was some other answer, or (d) "can't tell."

Table III
Proportion of Answers as a Function of Condition and Answer Category

	Answer Category			
	AI	Exception	Other	Can't Tell
Problem Form				
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

We can see from these data that Low Similarity tends to get the AI answer much more than High Similarity, and that within the Low Similarity condition the NEG subjects gave many more AI answers than the POS subjects. When we look at what happens with subjects who do *not* give the AI answer, the POS subjects tend to give more "can't tell" answers in the Lo Similarity condition than they do in the Hi Similarity condition. For Hi Similarity, the NEG subjects tend to say "other", whereas POS subjects are more likely to say that the object-in-question follows the exception object. We note that, with the possible exception of Pollock (1987), *no formal theory of defeasible reasoning predicts any effect about knowledge of an exception object upon the object-in-question, nor do any of them predict an effect for Similarity of an exception object with the object-in-question.* This seems a serious shortcoming of existing theories.

There is also a three-way interaction among answer category, specificity, and similarity in this study. First, when similarity is high, there was no significant impact of the specificity manipulation on the proportion of AI answers given (or for that matter, of the other answer categories). But when similarity is low, i.e., subjects are given the vanilla benchmark problems to solve, fewer AI answers are given under the specific form than under the nonspecific form. One interpretation of this result is that the impact of the extra information in the high-similarity case just outweighs any effect of specific or non-specific information about the exception. Second, we had predicted that the similarity and specificity manipulations would increase the proportion of "like the exceptional object" answers. Although the effect was in the right direction for the specific condition (.195 versus .090) and did approach significance, the increase in this particular answer was not reliable. Thirdly, the high proportion of "can't tell" answers under the high similarity/nonspecific case (.230) may be due to an artifact of the materials: perhaps the Hi Similarity subjects did not want to give the AI answer but there was no way in which they could easily say it was "like the exception" since no specific information about the exception was given. Instead, they drew their answers from the other information given in the high-similarity sentences added to the problem.

Returning now to the Robot/Human manipulation, let us take a look at the data:

Table IV
Proportion of Answers as a Function of Problem
and Problem Solver

Problem Solver:	Human				Robot			
	Benchmark Problem				Benchmark Problem			
	1	2	3	4	1	2	3	4
Answer Category								
AI	.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

Note (again) that Problem #3 is significantly different from the others. This problem is interesting because 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. Note also that when subjects imagine themselves as agents in the problems (the Human condition), Problem #3 is again different from the others, and that most of this difference is due to the high occurrences of “other” answers. In contrast, note that the difference in Problems #2 and #4 between the Human and the Robot condition is due our subjects wanting to have the Robot say “can’t tell” much more often than they require it of themselves. We enshrine this last feature as:

> *The Asimov Effect:* People believe robots should be cautious (saying they “Can’t tell”) in cases whey they themselves would be willing to give a definite answer.

Among the results of this experiment that we found interesting was the role that Hi vs. Lo Similarity played. Our view is that subjects interpret any "extra" information as possibly giving a reason why the exception object did not obey the default rule, and perhaps an indication that this reason applies to the object-in-question also. For example, in the Hi Similarity version of the experiment, a problem specified where computer manuals are typically found: subjects were told that both the IBM and the Mac manuals were being reviewed by support staff because new software had been purchased. This assertion could be interpreted as *a reason why* the exception object (the Mac manual) was not where manuals typically were, *and also could be construed as giving a reason* to believe the object-in-question (the IBM manual) might also violate the rule. We enshrine this view as:

> *Explanation-based Exceptions:* When the given information provides both a relevant explanation of why the exception-object violates the default rule and also provides a reason to believe that the object-in-question is similar enough in this respect that it will also violate the rule, then infer that the object *does* violate the rule.

So, it is not similarity alone, but rather the availability of information which explains why the exception *is* an exception, and which hints that the object-in-question might fall under that explanation.

Another interesting finding of this experiment was about Benchmark Problem #3, where it appeared that if an object violates one default rule, subjects will view it as being likely to violate other default rules as well. We enshrine this observation as:

> *Second-Order Default Reasoning:* If the available information is that the object-in-question violates other default rules, then infer that it will violate the present rule also.

(Others might prefer to call this the "Guilt by Past Association" rule, or maybe the "Bad Egg" principle ("once a bad egg, always a bad egg").

VI. Some Concluding Remarks

First we would like to emphasize the preliminary nature of these results. There are many further directions in which experimentation in this area could be taken, such as following up on the observation that subjects seem to be influenced by the specificity of the information they are given, by the amount of information available about the two objects, by the "similarity" that subjects perceive between the two objects, and by the number of previously-violated default rules. Each of these claims deserves more, and more careful, experimentation.

But turning our attention to the matters discussed in this paper, there are two types of conclusions that we would like to urge.

First type of conclusion: Unlike most other reasoning formalisms, nonmonotonic or default reasoning *is* "psychologistic" – that is, it *is* defined *only* as what people do in circumstances where they are engaged in "commonsense reasoning". It follows from this that such fundamental issues as "what are the good nonmonotonic inferences?" or "what counts as 'relevance' to a default rule?", etc., are only discoverable by *looking at people and how they behave*. It is *not* a formal exercise to be discovered by looking at mathematical systems, nor is it to be decided by such formal considerations as "simplicity" or "computability", etc.

Second type of conclusion: The results of the experiments reported here point to some particular considerations which seem to be critical to non-monontonic theories. First, we have some evidence that apparent similarity between the exception object and an object-in-question impacts whether or not the default rule would apply. The results for the basic default reasoning problems indicate that apparent similarity between two objects, and the specificity of information about how a rule is violated, impact the application of the default rule. How can relevance and specificity be worked into a logic of non-monotonic reasoning? One idea that we mentioned was called "explanation-based default reasoning", in the same sense used in the machine learning literature. That is, a reasoner attempts to explain why some default rule does not apply to one object and then sees if that explanation can fit another object. This would shift the investigation of non-monontonic reasoning away from a specification of a logic per se and towards the specification of some process that (perhaps deductively) determines whether an explanation constructed for one case fits a second case. We also drew attention to the notion of a “second-order default”, where an object is seen as violating many other default rules and is therefore seen as likely to violate the current rule also. The idea is that a default reasoning system should countenance “generally exceptionable objects”. And finally we drew attention to a feature that merits attention at some time: that people in general want to allow their artificial agents to draw fewer default conclusions then they themselves are willing to draw. This last is perhaps not an issue for default reasoning systems themselves to address, but is certainly a topic that needs to be addressed in the wider realm of intelligent artificial agents and their role in society.

Acknowledgements:

The authors would like to thank research assistants Siobhan Nearey and Matt Duncan, and the Department of Psychology at the University of Alberta for use of its research pool. They would also like to acknowledge the assistance of NSERC individual research grants, a joint NSERC equipment grant, and the University of Alberta Central Research Fund.

Bibliography

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

Baron , J. (1994) *Thinking and Deciding*. (NY: Cambridge Univ. Press).

Belnap, N. (1978) 'A Useful Four-Valued Logic' *Modern Uses of Multiple-Valued Logic* (ed.) J. Dunn & G. Epstein (Dordrecht: Reidel) pp. 8-37.

Cohen, L.J. (1981) 'Can Human Irrationality be Experimentally Demonstrated?' *Behavioral and Brain Sciences* 4: 317-370.

Evans, J. (1987) *Bias in Human Reasoning: Causes and Consequences.* (Hillsdale, NJ: Lawrence Erlbaum).

Gigerenzer, G. (1991) 'How to make Cognitive Illusions Disappear: Beyond "Heuristics and Biases" ' *European Review of Social Psychology* 2: 83-115.

Gigerenzer, G. (1994) 'Why the Distinction between Single-event Probabilities and Frequencies is Important for Psychology (and Vice Versa)' in G. Wright and P. Ayton (eds.) *Subjective Probability* (NY: John Wiley & Sons), pp. 129-161.

Gigerenzer, G. (1998) 'Ecological Intelligence: An Adaptation for Frequencies' in D. Cummins & C. Allen (eds.) *The Evolution of Mind* (NY: Oxford University Press), pp. 9-29.

Kahneman, D., P. Slovic, & A. Tversky (1982) *Judgment Under Uncertainty: Heuristics and Biases.* (Cambridge: Cambridge UP).

Lifschitz, V. (1989). 'Benchmark problems for formal nonmonotonic reasoning, version 2.00'. In M. Reinfrank, M., J. de Kleer, & M. Ginsberg (Eds.) *Nonmonotonic Reasoning*. (Berlin: Springer) 202-219.

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

McDermott, D. & 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.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. (San Mateo: Morgan Kaufmann).

Pelletier, F.J. & R. Elio (1997) 'What Should Default Reasoning Be, By Default?' *Computational Intelligence* 17: 165-187

Pollock, J. (1987) 'Defeasible Reasoning' *Cognitive Science* 11: 481-518.

Reiter, R. (1978) 'On Reasoning by Default' *Proceedings of TINLAP-2* (Univ. Illinois: Assoc. for Computational Linguistics) pp. 210-218.

Reiter, R. (1980) 'A Logic for Default Reasoning' *Artificial Intelligence* 13: 81-132.

Selman, B. & H. Kautz (1989). 'The Complexity of Model-Preference Default Theories'. In M. Reinfrank, M., J. de Kleer, & M. Ginsberg (Eds.) *Nonmonotonic Reasoning*. (Berlin: Springer) pp. 115-130.

Wason, P. (1968) 'Reasoning About a Rule' *Quarterly Jour. of Experimental Psychology* 20: 273-281.

Wason, P. & P. Johnson-Laird (1972) *The Psychology of Reasoning: Structure and Content* (Cambridge: Harvard Univ. Press).

Yates, J. F. (1990). *Judgment and Decision Making*. Englewood Cliffs: Prentice Hall.