

Influencing nonmonotonic reasoning by modifier strength manipulation

Kristien Dieussaert (kristien.dieussaert@psy.kuleuven.ac.be)

Department of Psychology, University of Leuven,
102 Tiensestraat, B-3000 Leuven, Belgium

Marilyn Ford (m.ford@griffith.edu.au)

School of Computing and Information Technology, Griffith University,
Nathan, Queensland, Australia, 4111

Leon Horsten (leon.horsten@hiw.kuleuven.ac.be)

Department of Philosophy, University of Leuven,
2 Kardinaal Mercierplein, B-3000 Leuven, Belgium

Abstract

Despite the current belief that much common sense reasoning is nonmonotonic in nature, research indicates that only a limited percentage of people are good at nonmonotonic reasoning. Good nonmonotonic reasoners recognize the logical strengths and weaknesses of some arguments. In the present study, we focus on differences in the probabilistic interpretation of the modifiers *typically* and *usually* and on the resulting differences in the strengths and weaknesses of arguments. We show that these implicit probabilistic strengths influence the reasoning process of good nonmonotonic reasoners.

Introduction

Most AI logicians and logic programmers, as well as philosophical logicians, ground their interest in nonmonotonic reasoning on the observation that common sense reasoning is largely nonmonotonic in nature. Ginsberg (1994, p.2), for example, states that: "... 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".

Pelletier and Elio (1997) argue that researchers like Ginsberg are right in grounding their research on human reasoning, but they also argue that AI researchers should bear the consequences of it.

As a first consequence, Pelletier and Elio (1997) argue for a more systematic study of human inferences. They plead against the use of the intuitions of a small group of AI researchers to decide on the acceptable answers to some 'benchmark' problems (e.g. Lifschitz, 1988). Within the AI community, at least some researchers are also supportive of this point of view (e.g. Schurz, 2001; Benferhat, Bonnefon, & Da Silva Neves, 2002).

The second consequence of grounding default reasoning research on human common sense reasoning is more severe. Contrary to deductive reasoning, where classical logic is considered to be the norm, there is no standard norm for

default reasoning. Since the long-term goal of most AI researchers is to simulate human reasoning, and since we do not have an objective theory of what rational default reasoning is, Pelletier and Elio (1997) argue that a psychological view of default reasoning should be adopted. Thus, the data that any default system should cover should be determined by the practices of ordinary people.

We agree with the importance of investigating human reasoning by controlled experimental research to gain more insight into the process of human reasoning. However, we also argue that it is important for AI not to develop a formal nonmonotonic logic or automated reasoning system on the basis of flawed reasoning.

Ford and Billington (2000) took the need for AI researchers to study human reasoning seriously and systematically investigated human nonmonotonic reasoning with abstract material. This way, participants could not rely on background knowledge, as they can do in daily life.

To clarify the discussion, we present an example of one problem, with a well-known Tweety-bird like structure:

Hittas are usually not waffs.

Penguins do not fly

All of the hittas are oxers.

All penguins are birds

Oxers are usually waffs.

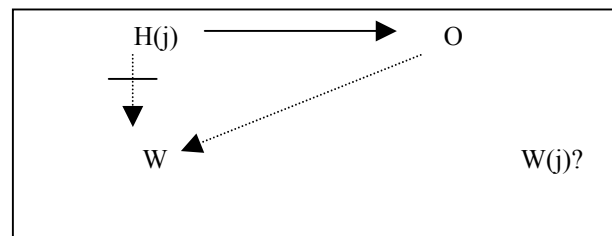
Birds fly

Jukk is a hitta.

Tweety is a Penguin

Is Jukk a waff?

Does Tweety fly?



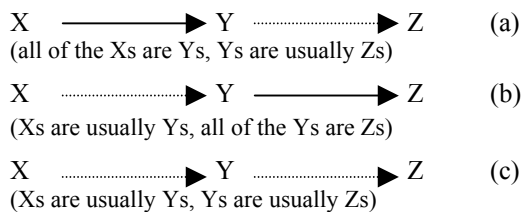
Ford and Billington (2000) summed up their main finding by presenting five negative (N) and three positive (P) factors

that influenced people's reasoning about nonmonotonic problems:

(N1) Most participants were not willing to draw a tentative conclusion when faced with conflict and with non-strict rules. (N2) Some participants weighed up the perceived number of relevant positive and negative paths, though the perceived paths were not paths. (N3) Some participants considered path length regardless of the ordering of rule types. Most participants preferred the shorter path to the longer path. (N4) Some participants gave weight to the presence of the universal quantifier, even when this was inappropriate. (N5) Some participants interpreted 'usually not' as evidence that 'some are' and thus gave preference to a positive conclusion. (See Hewson & Vogel, 1994, and Vogel & Tonhauser, 1996, for more evidence that many people have difficulty with nonmonotonic reasoning problems).

Besides these negative factors, Ford and Billington (2000) also extracted some positive factors from their experiment. (P1) Some participants recognized the relevance of the fact that if *all of the Xs are Ys* there might be *Ys that are not Xs*. (P2) Some participants recognized the relevance of the fact that if *all of the Ys are Zs* then *any Xs that are Ys are also Zs*. (P3) Some participants recognized that given a sentence *Xs are usually Ys* there are potentially many Ys that are not Xs.

People who recognize these positive points are able to see differences in the logical strength of arguments. Consider the following:



An appreciation of P2 allows people to recognize the strength of (b): for (b), it must be the case that Xs are usually Zs because the Xs that are Ys must also be Zs. In contrast, an appreciation of P1 and P3 allows people to see the weakness of (a) and (c), respectively: for (a) and (c), it might be that none of the Xs are Zs because it could be that the Ys that are not Xs are not Zs.

Ford (2004, In Press) argues that people who see the differences in the logical strength of arguments are more likely to give logically justifiable answers on nonmonotonic reasoning problems, since they rely on logically valid principles to form their answers. For example, with problems such as (1), they answer 'unlikely' more frequently and more strongly than people who do not see differences in the logical strength of arguments. They give this answer because of their recognition of the weakness of (a).

Note that these reasoners are not relying on a notion of 'specificity', where information stemming from a subclass

overrides information from a superclass; Ford and Billington's (2000) subjects did not articulate this notion of specificity as it is used in AI and the three P factors they identified make no mention of such specificity. The subjects instead rely on the logical strength of conflicting paths in an argument.

In this manuscript, we investigate further the nature of the logical strengths and weaknesses that reasoners who appreciate P2 and P3 use. We will argue for differences in the probabilistic interpretation of the modifiers *usually* and *typically* and consequent differences in the logical strengths and weaknesses of arguments. Given the results of Ford (2004, In Press), we would thus expect variations in conclusions given by reasoners who appreciate P2 and P3, with these reasoners giving more weight to the stronger side of an argument.

In a pilot experiment, we will extend a former study in which it was shown that researchers should be careful how to phrase their 'default relations' (Dieussaert, 2003). Researchers do not seem to make a distinction between sentences such as 'birds fly', 'birds normally fly', 'birds usually fly', 'birds typically fly' and so on. However, Dieussaert showed that the interpretation of these sentences, and the inferences yielded from them, differ greatly.

For the present study, we focus on the difference between *usually* and *typically*. We confirm the finding that *typically* is interpreted as indicating more instantiations of a type than *usually*. This implies that 'birds typically fly' represents a stronger relation than 'birds usually fly' since more instances of 'bird' are supposed to fly in the former case.

In a second experiment, we use this finding to show how the strengths and weaknesses of arguments can influence nonmonotonic reasoning. Reasoners who are shown to appreciate P2 and P3 are given problems with relations phrased with *typically* and *usually*. The data show clearly that the strength of arguments influences the nonmonotonic reasoning process for these subjects.

Pilot Experiment

In an earlier experiment (Dieussaert, 2003), the influence of phrasing on the interpretation of default sentences was shown. In a within subjects design, participants estimated the positive outcome of a sequence such as 'Hilo are typically waff. Jukk is a hilo. Is Jukk a waff?' significantly higher than for sequences like 'Brant are usually glent. Kerdo is a brant. Is Kerdo a glent?'

To obtain confirmatory evidence, we extended the earlier experiment.

Method

Participants

Ninety-nine first year students in Psychology at the University of Leuven, who had not taken a logic course, participated as a partial fulfillment of a course requirement.

Design

The design was completely within subjects. The dependent variable was the percentage entered per item.

Material and Procedure

Each participant received a booklet with written instructions and 18 items in randomized order. Each participant solved 18 problems: 9 positive items and 9 negative items. They solved this paper-and-pencil task individually and in a self-paced manner.

Here, we focus only on the difference between *typically* and *usually*, since these terms form the core of the main experiment.

The relevant material was:

- Nagdals are usually pirasos.
- Hittas are typically waffs
- Nilo are usually not riza
- Koki are typically not liri

The instructions for the positive items were as follows:

On each of the following pages a sentence will be presented. We ask you to mark how you interpret the underlined word within this sentence. To clarify the task, we give you an example.

The sentence: JY members are *normally* singers.

The question: Does the word ‘*normally*’ mean that:

Mark one or more answers:

0 A certain percentage of JY members have features that characterise singers. If so, given the sentence, what would you assume would be the approximate % (0-100) of JY members that have features that characterize singers.
.....% [further referred to as: **Feature**]

0 A certain percentage of time JY members are singers. If so, given the sentence, what would you assume would be the approximate % (0-100) of time JY members are singers.
.....% [further referred to as: **Time**]

0 A certain percentage of JY members are singers. If so, given the sentence, what would you assume would be the approximate % (0-100) of JY members that are singers.
.....% [further referred to as: **Number**]

For the negative items, the task were rephrased in a negative form e.g.: ...% (0-100) of JY members that are not singers.

Results

Table 1: Mean percentages given for Feature, Time, and Number (see Material).^a

Problem	Feature	Time	Number	Mean
Typically	91.8 (N=85)	65.0 (N=01)	92.3 (N=24)	91.8 [SD=11.7]
Usually	77.9 (N=31)	74.3 (N=17)	79.6 (N=70)	78.4 [SD=12.5]
Typically not	87.5 (N=73)	88.3 (N=04)	89.1 (N=36)	87.3 [SD=17.9]
Usually not	77.2 (N=27)	76.8 (N=38)	76.4 (N=52)	77.4 [SD=14.3]

^aThe number of responses (N) do not add up to 99 because participants were allowed to mark 1-3 answers. Some participants marked only the answers and did not enter a percentage.

Overall percentages entered for *typically* are higher than percentages entered for *usually* (91.8 vs. 78.4; $t(92) = 7.82$, $p < .00001$). Percentages entered for *typically not* are higher than percentages entered for *usually not* (87.3 vs. 77.4; $t(46) = 4.8$, $p < .00001$).

If we take into account only the single choices of participants (and remove items for which more than one answer was marked): Feature is the preferred category for *typically*, while Number is the preferred category for *usually*. A Sign test shows a higher number of Feature choices for *typically* (73) than for *usually* (17; Sign test, non-ties = 63, $Z = 6.3$, $p < .00001$). The same pattern is found for *typically not* (59) versus *usually not* (15; Sign test, non-ties = 70, $Z = 8.3$, $p < .00001$). A Sign test shows a higher number of Number choices for *usually* (55) than for *typically* (13; Sign test, non-ties = 56, $Z = 5.5$, $p < .00001$). A similar pattern is found for *usually not* (39) versus *typically not* (23; Sign test, non-ties = 33, $Z = 2.4$, $p < .00005$).

Discussion

This experiment confirms the results of Dieussaert (2003): *typically* and *usually* are interpreted somewhat differently. Most importantly for our purposes, having *typically* in a sentence is associated with higher percentages than having *usually*, with the former term thus being considered stronger.

Main Experiment

The pilot experiment provides confirmatory evidence for the stronger relation between two propositions A and B when they are connected in a default sentence with *typically* than with *usually*.

Having established this firmly, we can now investigate how the strength of this relation influences the nonmonotonic reasoning process. In this experiment, reasoners who appreciate Ford and Billington’s (2000) P2 and P3 are tested.

Method

Participants

Twenty-seven first year students in Psychology from the University of Leuven, who had not taken a logic course, participated as a partial fulfillment of a course requirement.

Design

The design was completely within subjects. The dependent variable was the score on a seven-point scale.

Material and Procedure

Each participant first received two critical questions (Ford, 2004) to see if they recognized Ford and Billington’s (2000) P2 and P3.

They were told that there were no time limits. The questions were:

1) Given the following two statements:

Mary's friends are usually Ann's friends.
 All of Ann's friends are Sue's friends.

Could it be the case that none of Mary's friends are Sue's friends? (Yes/No)

Given the following two statements:

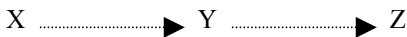
Jim's friends are usually Tom's friends.
 Tom's friends are usually Fred's friends.

Could it be the case that none of Jim's friends are Fred's friends? (Yes/No)

The first question contains the strong argument, with P2 relevant:



while the second question contains the weak argument, with P3 relevant:



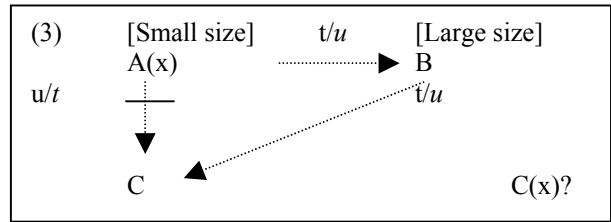
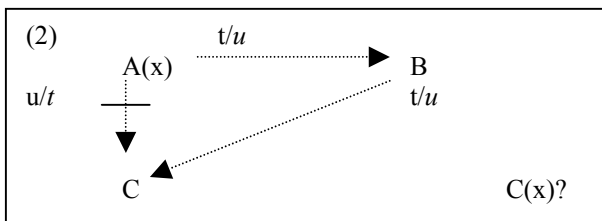
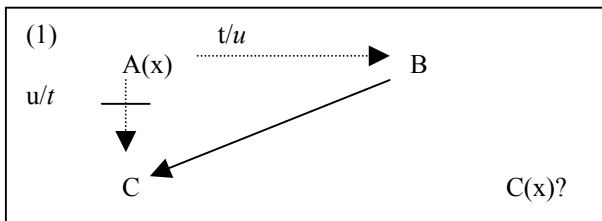
Only participants who answered the critical questions correctly (No on the first, Yes on the second) proceeded to the second part of the experiment, leaving 11 subjects. These participants received a booklet with written instructions and 18 problems (1 per page). Each participant gave their answer to each problem verbally and then indicated a likelihood estimation on a seven point scale.

On a scale of 1 to 7, estimate the likelihood of Z(x).

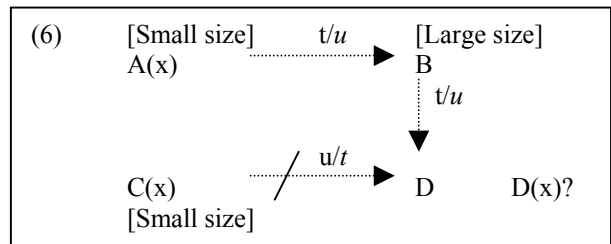
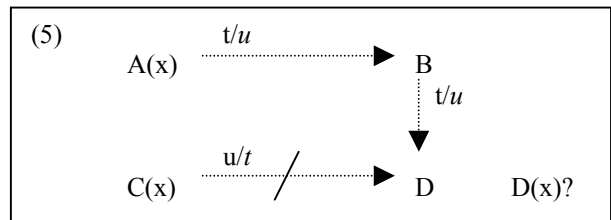
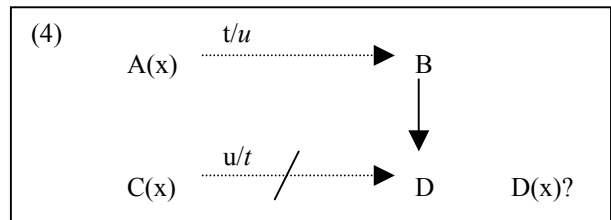


Problems in which the modifier *typically* (t) was used for the positive arguments and in which the modifier *usually* (u) was used for the negative arguments will be referred to as TU problems. Problems in which the modifier *usually* (u) was used for the positive arguments and in which the modifier *typically* (t) was used for the negative arguments will be referred to as UT problems.

Six differently structured problems can be distinguished. Participants received two examples of each of two versions of the following 3-argument structures, thus making 12 problems:



Participants also received one example of each of the two versions of the following 4-argument structures, making a further 6 problems:



It should perhaps be noted here that of the six structures studied here, the notion of specificity, if it were used, could only be applied to Problems 2 and 3, where information from a subclass (A) conflicts with information from its superclass (B). The good reasoners we are using, however, would be expected to use P2 and P3 to compare the strength of the conflicting arguments in all the problems.

Problems usually requiring 'Can't tell'

Problem 1 represents a strong positive versus a strong negative argument. The expected response when the modifier phrase is the same for both sides of the argument is thus around 4, meaning 'can't tell'. However, an additional manipulation was added.

In 1a, the positive non-strict relation was phrased with *typically* (t; As are typically Bs), while the negative non-strict relation was phrased with *usually* (u; As are usually not Cs). For problem 1b, the phrasing was vice versa (*usually* for the positive relation and *typically* for the negative one).

Problem 4 represents a similar problem to (1), but with four propositions involved. The expected response here is also 4, meaning ‘can’t tell’, when the modifiers are the same, but again TU and UT versions were given.

Given that *typically* is stronger than *usually*, the TU versions would be expected to result in a higher rating (more positive) than the UT versions.

Problems usually indicating ‘unlikely’

Problem 2 represents a weak positive versus a strong negative argument. The expected response when the modifier phrase is the same for both sides of the argument is thus lower than 4, meaning ‘unlikely’. However, the phrasing manipulation could be expected to have an additional influence on the final rating.

Problem 5 is similar to Problem 2, with a weak positive versus a strong negative argument, and with the phrasing manipulation expected to influence the final rating.

Problems 3 and 6 differ from 2 and 5, respectively, in that information is given on the relative subset/superset sizes to which the respective items belong: a small subset for A and a large one for B. It seems (Ford, In Press) that relative size information can sometimes help good reasoners in their reasoning.

For problems 2, 3, 5 and 6, the ratings for the TU versions would be expected to move higher, becoming more positive than would otherwise be expected. With the UT versions, the ratings would be expected to move lower, becoming even less positive than would otherwise be expected.

Results

Table 2: The mean likelihood ratings as a function of modifiers used in the positive and negative arguments. Standard deviations are given in square brackets.

(N = 11)	Problem	TU	UT
3-arg	(1)	4.3 [1.2]	3.7 [0.8]
	w/o relative size (2)	3.9 [1.1]	2.7 [0.8]
	with relative size (3)	4.2 [1.1]	2.5 [0.9]
4-arg	(4)	5.0 [1.7]	3.4 [1.2]
	w/o relative size (5)	4.2 [1.7]	2.5 [0.8]
	with relative size (6)	4.2 [1.3]	3.4 [1.4]

3-argument problems.

The mean likelihood rating was higher for TU problems than for UT problems (4.1 vs. 3.0; $F(1,10) = 17.5$, $MSE = 1.2$, $p < .005$). Planned comparisons showed that TU ratings are higher than UT ratings for Problem 1 (4.3 vs. 3.7; $F(1,10) = 7.7$, $MSE = .2$, $p < .05$), for Problem 2 (3.9 vs. 2.7; $F(1,10) = 10.1$, $MSE = .7$, $p < .01$), and for Problem 3 (4.2 vs. 2.5; $F(1,10) = 14.4$, $MSE = 1.1$, $p < .005$).

No difference between the problems was observed ($p = .08$). However, an interaction between problem and modifier was observed ($F(2,20) = 4.4$, $MSE = .4$, $p < .05$). A planned comparison shows only a significant difference for UT between Problem 1 and 2 (3.7 vs. 2.7; $F(1,10) = 6.9$, $MSE = .8$, $p < .05$) and between Problem 1 and 3 (3.7 vs. 2.5; $F(1,10) = 31.0$, $MSE = .3$, $p < .0005$). This difference is due to the particularly low ratings of UT Problem 2 and 3. Notice, too, that relative size information did not influence the ratings.

4-argument problems.

A similar pattern is found for 4-argument problems. The mean likelihood rating was higher for TU problems than for UT problems (4.5 vs. 3.1; $F(1,10) = 7.0$, $MSE = 4.5$, $p < .05$). Planned comparisons showed that TU ratings are higher than UT ratings for Problem 4 (5.0 vs. 3.4; $F(1,10) = 5.2$, $MSE = 2.8$, $p < .05$), for Problem 5 (4.2 vs. 2.5; $F(1,10) = 6.8$, $MSE = 2.4$, $p < .05$), but not for Problem 6 (4.2 vs. 3.4; $p = .2$).

No difference between the problems was observed ($p = 0.5$). No interaction between problem and modifier was observed.

Discussion

This study was set up to gain more insight into the role that modifiers of non-strict relations play in the nonmonotonic reasoning process. Generally, researchers do not pay much attention to the specific wording of default expressions. We showed in a pilot experiment that this neglect is undeserved: default expressions vary in interpretation. However, only if this interpretation also affects the nonmonotonic reasoning process does the topic become particularly noteworthy for AI researchers and philosophical logicians doing research on nonmonotonic reasoning.

In the main experiment we showed that the use of different modifiers in non-strict relations does indeed lead to a variation in nonmonotonic reasoning, more precisely in likelihood ratings on nonmonotonic reasoning problems. We presented reasoners who appreciate P2 and P3, with problems of two kinds: problems with structures where we would normally expect them to give a ‘can’t tell’ answer and problems where we normally expect them to give an ‘unlikely’ answer. So, if participants bore only the structure of the problem in mind, we would expect a ‘can’t tell’ answer for Problems 1 and 4, and an ‘unlikely’ answer for Problem 2-3 and 5-6, despite the modifier manipulation. However, if the reasoning process was influenced by the modifier used, we would see a shift in answers, depending on the specific modifier used to express the positive and negative non-strict relations.

The data show clearly that reasoners who appreciate P2 and P3 are influenced by the modifier used. Ratings on TU problems differ significantly from ratings on UT problems. With the ‘can’t tell’ Problem 1, we observed ratings staying close to the ‘can’t tell’ rating, although a positive shift was noted for TU problems, while a slightly negative shift was noted for UT problems, resulting in an overall difference.

With the 'can't tell' Problem 4, the pattern was more extremely pronounced, with a large positive shift for TU problems and a large negative shift for UT problems.

The TU versions of problems 2-3 and 5-6 are lifted up to a 'can't tell' level, while UT versions receive an 'unlikely' rating. While adding relative size information has been shown to sometimes help good nonmonotonic reasoners (Ford, In Press), it did not affect the reasoning process in our experiment, possibly because these subjects did not need this help.

It is clear that although the matching TU and UT versions of problems have the same structure, they are not considered as being equivalent. The modifier *typically* or *typically not* makes a non-strict relation stronger compared with its counterpart *usually* or *usually not*.

It is clear that people who show an appreciation of P2 and P3 and who solve nonmonotonic problems by comparing the logical strength of conflicting arguments, rather than by using a notion of specificity, also use the strength of modifiers to guide their reasoning. Just as it is rational to take the logical strength of conflicting arguments into account, rather than using a notion of specificity, so too it is rational to take into account modifier strength in conflicting arguments.

Conclusion

Ford (2004, In Press) has shown that good reasoners use the logical strength of different sides of an argument to guide their reasoning. The present study adds credence to this effect of weighing up the strengths of the different sides of an argument. The study shows that different modifiers can differentially weaken or strengthen an argument and that they thereby influence reasoning.

Acknowledgments

This research was made possible by the financial support of the Fund for Scientific Research Flanders (project G.0239.02: K. Dieussaert and L. Horsten).

References

- Benferhat, S., Bonnefon, J. F., & Da Silva Neves, R. M. (2002, July). *An overview of possibilistic handling of default reasoning: Applications and empirical studies*. Paper presented at the 1st Salzburg Workshop on Paradigms of Cognition (SWPC 1/2002), 'Nonmonotonic and uncertain reasoning in the focus of competing paradigms of cognition', Salzburg, Austria.
- Dieussaert, K. (2003). Do typical birds usually fly normally?. In A. Markman & L. Barsalou (Eds.). *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Boston, MA: Cognitive Science Society, Inc.
- Ford, M. (2004). System LS: A three-tiered nonmonotonic reasoning system. *Computational Intelligence*, 20 (1), 89-108.

Ford, M. (In Press). Human nonmonotonic reasoning: The importance of seeing the logical strength of arguments. *Synthese*.

Ford, M., & Billington, D. (2000). Strategies in human nonmonotonic reasoning. *Computational Intelligence*, 16 (3), 446-468.

Ginsberg, M. L. (1994). AI and nonmonotonic reasoning. In D. M. Gabbay, C. J. Hogger, & J. A. Robinson (Eds.), *Handbook of logic in artificial intelligence and logic programming. Vol. 3: Nonmonotonic reasoning and uncertain reasoning*. Oxford: Clarendon Press.

Hewson, C. and Vogel, C. (1994). Psychological evidence for assumptions of path-based inheritance reasoning. *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Atlanta, Georgia.

Lifschitz, V. (1988). Benchmark problems for formal nonmonotonic reasoning, version 2.00. In J. Siekmann (Series ed.) & M. Reinfrank, J. de Kleer, M.L. Ginsberg, & E. Sandewall (Vol. Eds.), *Lecture notes in Artificial Intelligence. Nonmonotonic reasoning*. Berlin : Springer-Verlag.

Pelletier, F. J., & Elio, R. (1997). What should default reasoning be, by default? *Computational Intelligence*, 13, 165-187.

Schurz, G. (2001). What is 'normal'? An evolution-theoretic foundation for normic laws and their relation to statistical normality, *Philosophy of Science*, 68 (4), 476-497.

Vogel, C. and Tonhauser, J. (1996). Psychological constraints on plausible default inheritance reasoning. In Aiello and Shapiro (Eds.), *Proceedings of the 5th International Conference on Principles and Practice of Knowledge Representation, KR'96*. Cambridge, Mass.: Morgan Kaufmann. 608-19.