Reasoning Biases, Non-Monotonic Logics and Belief Revision

Dutilh Novaes, Catarina; Veluwenkamp, Herman

Published in:
Theoria

DOI:
10.1111/theo.12108

Publication date:
2017

THEORIA

THEORIA, 2017, **83**, 29–52
doi:10.1111/theo.12108

Reasoning Biases, Non-Monotonic Logics and Belief Revision

by

CATARINA DUTILH NOVAES

and

HERMAN VELUWENKAMP

University of Groningen

Abstract: A range of formal models of human reasoning have been proposed in a number of fields such as philosophy, logic, artificial intelligence, computer science, psychology, cognitive science, etc.: various logics (epistemic logics; non-monotonic logics), probabilistic systems (most notably, but not exclusively, Bayesian probability theory), belief revision systems, neural networks, among others. Now, it seems reasonable to require that formal models of human reasoning be (minimally) empirically adequate if they are to be viewed as models of the phenomena in question. How are formal models of human reasoning typically put to empirical test? One way to do so is to isolate a number of key principles of the system, and design experiments to gauge the extent to which participants do or do not follow them in reasoning tasks. Another way is to take relevant existing results and check whether a particular formal model predicts these results. The present investigation provides an illustration of the second kind of empirical testing by comparing two formal models for reasoning – namely the non-monotonic logic known as preferential logic; and a particular version of belief revision theories, screened belief revision – against the reasoning phenomenon known as belief bias in the psychology of reasoning literature: human reasoners typically seek to maintain the beliefs they already hold, and conversely to reject contradicting incoming information. The conclusion of our analysis will be that screened belief revision is more empirically adequate with respect to belief bias than preferential logic and non-monotonic logics in general, as what participants seem to be doing is above all a form of belief management on the basis of background knowledge. The upshot is thus that, while it may offer valuable insights into the nature of human reasoning, preferential logic (and non-monotonic logics in general) is ultimately inadequate as a formal model of the phenomena in question.

Keywords: preferential logics, belief bias, belief revision, formal models of reasoning

1. Introduction

A RANGE OF FORMAL MODELS of human reasoning have been proposed in a number of fields such as philosophy, logic, artificial intelligence, computer science, psychology, cognitive science, etc.: various logics (epistemic logics; non-monotonic logics), probabilistic systems (most notably, but not exclusively, Bayesian probability theory), belief revision systems, neural networks, among others. Insofar as they are *models*, each of these systems will inevitably display different levels of idealization and simplification with respect to the "messy" phenomena they

represent. Moreover, some of these models are presented as having a normative rather than descriptive character, which in turn may be seen as implying that a significant mismatch between actual patterns in human reasoning and the models in question is not an issue.[1] Here we focus on models that present themselves as at least partially descriptive (though they may also be seen as having normative import).

However, even if highly idealized, it seems reasonable to require that formal models of human reasoning be (minimally) empirically adequate if they are to be viewed as models of the phenomena in question. In other words, testing these models empirically, as it were, seems like a perfectly reasonable methodological desideratum. Naturally, the issue of the empirical adequacy of scientific models/theories is a very general question within the philosophy of science, which has been and continues to be extensively discussed (a seminal text in these debates is Van Fraassen, 1980). Indeed, the empirical adequacy of formal models of human reasoning is simply a specific instantiation of a much more general scientific question. And yet, while many of the theorists involved in formal modelling of human reasoning are well aware of the significance of at least some level of empirical adequacy for their models (in particular in cognitive science, where empirically-grounded formal modelling is an important trend), this is arguably not yet sufficiently recognized by a number of modellers in different fields.

How are formal models of human reasoning typically put to empirical test? There is nothing particularly exotic about it when compared to how scientific theories in general are empirically tested. One way to do so is to isolate a number of key principles of the system, and design experiments to gauge the extent to which participants do or do not follow these principles in reasoning tasks. Another way is to take relevant existing results from empirical studies of human cognition (e.g., in psychology and cognitive science), preferably robust results having emerged from a large number of studies, and evaluate whether a particular formal model predicts these results. If this is the case, then the model can be said to pass this particular empirical test (which of course does not mean that the model is once and for all empirically confirmed). If, however, the model makes predictions that clash with the robust empirical data in question, then it seems that the model must be viewed as having limited empirical adequacy. This does not mean that the model thereby becomes "useless", as it may still be a valuable tool to investigate different aspects of the phenomenon in question (with respect to which it is empirically adequate), but it does mean that its level of empirical adequacy is lower than might have been thought at first.

1 It can be argued, however, that even normative models should be to some extent accountable towards the empirical phenomena in question, but we will leave this issue aside for now.

This kind of empirical testing can also be used to compare two different formal models and determine which one is better supported by the data in question; naturally, *ceteris paribus*, the model with a higher degree of empirical adequacy is to be preferred. Furthermore, a comparison of the formal properties of the two systems which give rise to the different predictions may also tell us something about the phenomenon which we may not have noticed before.

The present investigation provides an illustration of this second kind of empirical testing by comparing two formal models for reasoning – namely the non-monotonic logic known as preferential logic and a particular version of belief revision theories, screened belief revision – against the reasoning phenomenon known as belief bias in the psychology of reasoning literature. Belief bias is the tendency that reasoners display to let the (un)believability of the conclusion influence their judgement of the (in)validity of an argument. It has been described as "perhaps the best known and most widely accepted notion of inferential error to come out of the literature on human reasoning" (Evans, 1989, p. 41), and is related to a number of other empirically observed cognitive phenomena which all point in the same direction: human reasoners typically seek to maintain the beliefs they already hold, and conversely to reject contradicting incoming information, thus typically letting background knowledge play an important role in reasoning processes. This can be described as a tendency towards doxastic conservativeness. The conclusion of our analysis will be that screened belief revision is more empirically adequate with respect to belief bias and doxastic conservativeness than preferential logic and non-monotonic logics in general.

Prima facie, non-monotonic logics – and preferential logic in particular, given the key notion of preferred models – seem like a promising formal explanans for the concept of doxastic conservativeness, and for the general idea of bringing background knowledge (prior beliefs) to bear in reasoning tasks. Non-monotonic logics have figured prominently in the recent work of Stenning and van Lambalgen (2008; 2010). They adopt their preferred non-monotonic framework, namely closed-world reasoning, as a theoretical basis both to re-interpret previous results in a number of well-known reasoning experiments – Wason selection task, suppression task, etc. – and for the formulation and interpretation of new experiments. Despite the comprehensiveness of their investigations, Stenning and van Lambalgen did not discuss belief bias and related phenomena.

Besides the work of Stenning and van Lambalgen, a number of studies (such as Benferhat et al., 2005; Pfeifer and Kleiter, 2005) have taken as their starting point some of the basic principles of well-known non-monotonic systems, and formulated experiments to test the extent to which participants do reason according to these principles (thus exemplifying the first approach to empirical testing described above). The results have for the most part indicated similarities between

the reasoning behaviour of participants at experiments and these basic principles. However, there have only been a handful of such studies, and so the question of the empirical plausibility and adequacy of non-monotonic logics as accurate descriptions of human reasoning requires further scrutiny at this point.

Empirical adequacy of the preferential framework against the belief bias results, if obtained, would lend further support to the claim that non-monotonic logics may represent a plausible descriptive model of human reasoning (as defended by Stenning and van Lambalgen). In effect, many instances in which participants seem to be performing deductive reasoning incorrectly can also be explained as instances of participants in fact correctly performing defeasible reasoning instead. However, the comparison with the data will also highlight the limitations of this framework. Indeed, a large chunk of the experimental results to be discussed cannot be straightforwardly explained from the point of view of preferential logic (or other supraclassical non-monotonic logics for that matter), thus establishing the limitations of this framework as a descriptive model of human reasoning.

The alternative framework of belief revision theory (in one of its variants, namely screened belief revision), which is known to be closely related to non-monotonic logics (Makinson and Gärdenfors, 1991), seems to allow for a better fit with the belief bias experimental data as a whole, including the data that cannot be accounted for within the preferential framework, thus outlining what may be "missing" in the latter.[2] In other words, the predictions of screened belief revision are essentially borne out by these empirical results, which thus means that screened belief revision is more empirically adequate than preferential logic with respect to these data. What this also suggests is that, in the case of the belief bias experiments at least, what participants are in fact performing is neither indefeasible deductive reasoning nor defeasible reasoning; instead, they are engaging in what could be described as belief management – at each incoming piece of information, determining whether it should be incorporated to one's belief set or not on the basis of background knowledge and the reliability of the new information.[3]

The article proceeds as follows. We first present the experimental data on belief bias and related phenomena, which will later be used to test the empirical

2 The classical AGM belief revision framework is formally equivalent to preferential logic (the KLM system), and this is why we need to turn to a different variant of belief revision theory in order to account for the data that preferential logic cannot account for. What is peculiar about screened belief revision is precisely that background knowledge and prior beliefs play a more central role than in AGM.

3 The observation that there may be a discrepancy between what experimenters aim at and how participants in fact interpret the task at hand has been made before by Oaksford and Chater (1991) and Stenning and van Lambalgen (2008; 2010); the latter emphasize the importance of the process of reasoning *to* an interpretation of the task materials, not only *from* an interpretation.

adequacy of preferential logic and screened belief revision. In the second part we introduce preferential logic and apply the framework specifically to the experimental results discussed in the first section. The outcome is that, while this logic predicts some of these results, it fails to predict a significant subgroup of them. In the third part, we discuss a specific version of belief-revision theory (screened belief revision), and show that its predictions are very much aligned with the belief bias data. We then offer some philosophical observations in the fourth part. The overall conclusion is thus that the formal treatment of the notion of most preferred models in preferential logic ultimately fails as a formal explanans for the phenomenon of doxastic conservativeness (as exemplified by the belief bias results) in human reasoning, whereas screened belief revision seems to fare better in this respect.

2. Experimental data

As is well known, one of the key concepts to have emerged from the psychological literature on reasoning and decision-making is the concept of *cognitive biases*. It figures prominently in Tversky and Kahneman's (1974) "heuristics and biases" research programme, as well as in the literature on reasoning stemming from the pioneer work of P. Wason in the 1960s. With respect to reasoning in particular, biases were initially conceptualized as systematic deviations from the canons dictated by classical logic (or some other "traditional" normative system). However, as the status of classical logic as the only legitimate normative system for human reasoning began to be questioned (Oaksford and Chater, 1991), the association between the concept of biases and the notion of reasoning "mistakes" began to be questioned as well (although the strong normative conception is still widespread). As described by Klayman (1995), a bias can be understood in (at least) three ways:

- A tendency or inclination (neutral)
- A flawed reasoning tendency (negative)
- Bounded rationality: people may deviate systematically from theoretical standards, but may still be behaving optimally when broader concerns and limitations are taken into account.

For the present investigation, the main point is that what counts as a faulty response from the point of view of deductive, monotonic reasoning, purportedly betraying the effect of reasoning biases, may just as well count as an adequate response from the point of view of different normative standards. (This general point has been made before, in particular by Oaksford and Chater with respect to Bayesian normative standards.) However, adopting different normative standards

– and accordingly, different formal frameworks to guide the investigation – does not mean that the very notion of reasoning errors will fade away: reasoners may still make reasoning mistakes even if what they are trying to do is to perform, say, defeasible reasoning.

In the psychological literature, a number of phenomena have been identified which all seem to point in the same direction: a tendency to reason towards maintaining the beliefs we already hold, which can be described as a tendency towards *doxastic conservativeness*. This general tendency manifests itself in several ways, and one general term that is often used to refer to a number of related phenomena is "confirmation bias" (Klayman, 1995; Nickerson, 1998).

In a similar vein, Kahneman (2011) coined the acronym "WYSIATI" to refer to the principle "what you see is all there is", which manifests itself in the form of several of the "biases" he discusses throughout the book. Here the idea is the privileged status accorded to positive available information – prior beliefs, background knowledge – when reasoning, leading to a tendency of "jumping to conclusions on the basis of limited evidence" (Kahneman, 2011, p. 86). In effect, this is a rather accurate description of what non-monotonic reasoning is about: "jumping to conclusions" on the basis of partial information, which is of course all we have to go by on numerous occasions. Indeed, even those (such as Kahneman) who see it as giving rise to biases recognize that this tendency is by and large (though not always) a reliable guide for action.

> WYSIATI facilitates the achievement of coherence and of the cognitive ease that causes us to accept a statement as true. It explains … how we are able to make sense of partial information in a complex world. Much of the time, the coherent story we put together is close enough to reality to support reasonable action. However, I will also invoke WYSIATI to help explain a long and diverse list of biases of judgment and choice … (Kahneman, 2011, p. 87)

In the reasoning literature more specifically, one (extensively investigated) tendency that participants in experiments seem to have is to endorse arguments whose conclusions they believe, and likewise to reject arguments whose conclusions they disbelieve, irrespective of their actual validity; this is again a manifestation of a general inclination towards doxastic conservativeness in human reasoners. The term commonly used to refer to this tendency is "belief bias". Stanovich (2003, p. 292) also speaks of a "fundamental computational bias": "the tendency to automatically bring prior knowledge to bear when solving problems". Stanovich's fundamental computational bias as such does not entail doxastic conservativeness – i.e., it is not *sufficient* for doxastic conservativeness – but it is clearly connected to the inclination to maintain one's previous beliefs. So all in all, it seems that, when reasoning, we spontaneously bring in prior beliefs and background knowledge, and then reason towards a story that maximizes accommodation of these

Table 1. *Arguments used in the study*

Valid-believable	Valid-unbelievable	Invalid-believable	Invalid-unbelievable
No police dogs are vicious.	No nutritional things are inexpensive.	No addictive things are inexpensive.	No millionaires are hard workers.
Some highly trained dogs are vicious.	Some vitamin tablets are inexpensive.	Some cigarettes are inexpensive.	Some rich people are hard workers.
Therefore, some highly trained dogs are not police dogs.	Therefore, some vitamin tablets are not nutritional.	Therefore, some addictive things are not cigarettes.	Therefore, some millionaires are not rich people.

prior beliefs to the data of the problem (and vice versa). Coherence is usually achieved by maintaining entrenched prior beliefs as much as possible.[4]

So let us look at some of the experimental data supporting these claims. In Evans et al. (1983),[5] participants were presented with fully formulated syllogistic arguments (embedded in longer texts, so as to reduce artificiality) and then asked to evaluate whether a given conclusion could be "logically deduced" (or if it "necessarily follows") from the information contained in the text. The syllogistic arguments presented were of four types: valid arguments (valid according to traditional syllogistic, that is) with believable conclusions; valid arguments with unbelievable conclusions; invalid arguments with believable conclusions; invalid arguments with unbelievable conclusions.[6] Table 1 displays some examples of the arguments used in the study. (Notice that the two valid arguments have the same "logical form", and the same holds for the two invalid ones.) Table 2 shows the percentages of arguments whose conclusions were said to "follow necessarily" from the premises by the participants.

It is immediately apparent that there are some interesting correlations between the participants' responses and the (un)believability of the conclusions: arguments with believable conclusions were much more often endorsed than those with

4 Naturally, this is a view also familiar from the philosophical literature, most notably defended by Quine, for example in Quine (1955).

5 One may quibble with the fact that this study is already 30 years old. However, the belief bias effect has been replicated several times in a large number of studies since, and the general pattern described in this study has been confirmed time and again.

6 Typically, belief bias experiments include a preliminary, independent step of believability evaluation of the conclusions taken in isolation.

Table 2. *Percentages of arguments whose conclusion were said to "follow necessarily"*

	Believable conclusion	Unbelievable conclusion
Valid	89	56
Invalid	71	10

unbelievable conclusions, both for valid and for invalid arguments. Validity also had an effect, as valid arguments were more often endorsed than invalid ones in each of the categories (believable vs. unbelievable conclusions). But what is perhaps most striking is that *in*valid arguments with believable conclusions were more often endorsed (71%) than valid arguments with *un*believable conclusions (56%). Clearly, external prior beliefs and background knowledge were brought in when participants were evaluating these arguments, even though the instructions referred specifically to the concept of conclusions being "logically deduced", which pertains exclusively to the relation between premises and conclusions, not to their truth or believability.[7] In effect, arguments whose conclusions confirm the reasoner's prior beliefs were deemed correct much more often than those whose conclusions went against the reasoner's prior beliefs.

To further probe the effect of conclusion believability, Stanovich and collaborators (Sá et al., 1999) designed an experiment including arguments whose conclusions would be neither believable nor unbelievable, given that they were composed of invented words. They started by giving participants an invalid syllogistic argument (again, invalid according to traditional syllogistic) with a believable conclusion:

All living things need water.
Roses need water.
Therefore, roses are living things.

As could have been anticipated, only 32% of the participants identified this argument as invalid. Subsequently, the same group of participants was given a little

7 If an argument is valid *and* has true premises, then it is described as a sound argument; sound arguments are thus a subclass of valid arguments, as the truth of premises is not required for validity. It is of course questionable whether untrained participants would possess the appropriate concept of "logical deduction" or "validity" so as to understand what exactly was expected of them in the task. This is an important methodological objection to these studies, but it does not invalidate the observation that participants endorse arguments with believable conclusions much more readily than those with unbelievable conclusions. Indeed, it underscores the fact that the initial stage where a participant preprocesses the input requires further scrutiny, as also pointed out by Stenning and van Lambalgen.

scenario about another planet with different animals, and the following syllogistic argument of the same "logical form" as the previous one (and thus invalid):

All animals of the hudon class are ferocious.
Wampets are ferocious.
Therefore, wampets are animals of the hudon class.

This time, 78% of the very same participants deemed this argument to be invalid, as believability of conclusion was not a factor this time around. (Presumably, participants have no prior beliefs concerning made-up words such as "wampets" and "hudon".) Stanovich (2003) presents these results as offering strong support to his ascription of a "fundamental computational bias" to human reasoners. Moreover, notice that the instructions in both cases were identical, so the general worry that participants do not understand exactly what it means for a conclusion to follow logically from the premises seems somewhat attenuated by the observation that in this case, the vast majority apparently "knew what to do" (from the point of view of what the experimenter expected them to do).

One may object that requiring participants to evaluate fully formulated arguments does not really capture their reasoning behaviour in real-life situations. In particular, we should mainly be interested in how they draw conclusions themselves, rather than in how they evaluate previously drawn conclusions. In effect, experiments with conclusion-production tasks have also been conducted, and again the belief bias pattern emerges quite robustly. For example, in Oakhill and Johnson-Laird (1985), participants were given pairs of syllogistic premises and asked to choose one option from a list of five options, which included conclusions following necessarily from the premises, other sentences with the same terms, as well as the option "no valid conclusion" (among those listed).[8] One of the pairs presented to participants was:

Some of the actresses are not beautiful.
All of the women are beautiful.

According to syllogistic, this pair of premises produces a necessary conclusion, as it is an instance of the valid syllogistic mood "Some *A* are not *B*. All *C* are *B*. Therefore, some *A* are not *C*". Thus, the correct response from the point of view of classical deductive logic/syllogistic is "Some of the actresses are not women". This is, however, a highly unbelievable conclusion (in fact the authors describe it as "definitionally false"), clashing violently with the reasoner's background

8 It may still be objected that this is yet not a case of conclusion *production*, properly speaking, but given the requirement of a controlled experimental setting, this is probably the best that can be hoped for.

knowledge. And indeed, participants' responses indicate yet again unwillingness to revise their prior beliefs concerning actresses not being women: only 38% of the participants chose the syllogistically correct response, "Some of the actresses are not women" as the conclusion; 46% of them said there was no valid conclusion to be drawn (16% gave other responses). This result illustrates the pattern of refusing to draw a conclusion that does not accord with prior belief, even if the reasoner is instructed to focus on the notion of logical validity.

Participants were also given pairs of premises from which no conclusion can be drawn according to classical syllogistic (i.e., a sentence where a term from each of the premises is connected to the other term in a categorical sentence of the form "All *A* is *B*", "Some *A* is *B*", "No *A* is *B*" or "Some *A* is not *B*", and which follows necessarily from the premises). One example is:

Some of the women are not beautiful.
All of the beautiful people are actresses.

The premises instantiate the forms "Some *A* are not *B*" and "All *B* are *C*" from which no syllogistic conclusion can be drawn. This time, only 17% of participants opted for the syllogistically correct "No valid conclusion" response (i.e., none of the alternatives provided follows necessarily from the premises). Forty-six per cent opted for "Some of the women are not actresses" as a conclusion following from the premises, presumably because this is a statement with a high degree of believability involving the terms in the premises. (There was a residue 37% of other responses.) This result illustrates the pattern of "jumping to a conclusion" that is not a deductive conclusion from the premises but is highly believable.

So there are two patterns of deviation from the "correct responses" from the point of view of traditional deductive logic (based on the concept of necessary truth-preservation):

Undergeneration: There is a conclusion that does follow necessarily from the given premises, but which reasoners refuse to draw if it clashes with their prior beliefs.

Overgeneration: There is a "conclusion" that does not follow necessarily from the given premises, but which reasoners readily draw if it also has a high degree of believability.

The experimental data suggest that overgeneration occurs more frequently than undergeneration. In the first study cited here, for example, in 71% of the cases participants gave the "wrong" reply when the argument was invalid with a believable conclusion. By contrast, in only 44% of the cases did participants incorrectly deem a deductively valid argument with unbelievable conclusion to be invalid. As we will see, preferential logic can successfully

account for the more robust phenomenon of overgeneration, but they fail to account for the less robust but nevertheless clearly pervasive phenomenon of undergeneration.

3. Preferential logic

Preferential logic was first introduced in Shoham (1987), where a semantic framework for nonmonotonic logics was proposed with the goal of providing a unified analysis for the several non-monotonic logics then available in the literature. The key idea is the concept of *most preferred models*,[9] which allows for the definition of a preferential consequence relation: *A is a preferential consequence of* Γ *iff A is true in all of the most preferred models of* Γ. This definition contrasts with the classical definition of logical consequence, which requires A to be true in absolutely *all* models of Γ, not only the most preferred ones. The framework can accommodate different preference criteria, thus generating different non-monotonic logics by associating a "standard" monotonic logic with different preference relations over models. (There are restrictions on what counts as a legitimate preference relation for this purpose.)

Here is the general idea. Take a monotonic logic L; since L is monotonic, the following property holds: for all A, B and C in L, if $A \Rightarrow C$, then also $A, B \Rightarrow C$. Then define a strict partial order[10] $\prec$ on the class of models M of L: $M_1 \prec M_2$ means that M_1 is preferred over M_2. $L_\prec$ is the non-monotonic logic generated from L and $\prec$.

Definition. Let M be a model and Γ a finite set of formulae. Then M *preferentially satisfies* Γ ($M \models_\prec \Gamma$) iff M is a model of Γ ($M \models \Gamma$), and there is no other model M' such that $M' \prec M$ and $M' \models \Gamma$. M is a *most preferred (or minimal) model* of Γ.[11]

Definition. A is a *preferential consequence* of Γ ($\Gamma \Rightarrow_\prec A$) iff for any M, if $M \models_\prec \Gamma$, then $M \models A$. That is, the class of models of A (preferred or otherwise) is a *superset* of the class of preferred models of Γ.

It is easy to see that $\Rightarrow_\prec$ is a non-monotonic consequence relation. If an arbitrary C is added to the antecedent, now the consequence obtains only if B holds in all most preferred models of *A and C* together. But it may well be that $\{A, C\}$

9 A model of a set of sentences is an interpretation in which all sentences in the set are true.

10 A strict partial order is a binary relation that is irreflexive, transitive and asymmetric. Non-strict partial orders (reflexive, transitive and anti-symmetric) or, more generally, pre-orders (reflexive and transitive) are also used to generate different non-monotonic logics.

11 The qualification "most" can be dropped for convenience of expression.

has preferred models that are not preferred models of A alone; in fact, the two classes of preferred models may even be disjoint.[12] And it may well happen that in at least one preferred model of $\{A, C\}$ that is not a preferred model of A alone, B does not hold: so the consequence relation no longer holds with the addition of C to the antecedent.

One of the advantages of the preferential framework is that it allows for an illuminating characterization of the divide between defeasible and indefeasible reasoning. Indefeasible reasoning would be a limit case of defeasible reasoning, namely the case where *all* models are preferred models – in other words, where the preferential order has all models as minimal and is thus no longer an *order*, so to speak. The idea is that for indefeasible reasoning, each and every model is equally "normal". Indeed, the difference between monotonic reasoning and (non-monotonic) preferential reasoning is that the former requires that the reasoner takes into account *each and every model* of the premises, whereas the latter restricts the requirement to a subclass of the models of the premises, namely the most preferred ones.

When he proposed the framework of preferential logic, Shoham had the explicit concern of capturing an actual feature of the cognitive makeup of human reasoners. As described in the classic Kraus et al. (1990) (usually referred to as KLM):

> He [Shoham] suggested models that may be described as a set of worlds equipped with a preference relation: the preference relation is a partial order and a world v is preferable, in the eyes of the reasoner, to some other world w if he considers v to be more normal than w. One would then, in the model, on the basis of a proposition α, conclude, defeasibly, that a proposition β is true if all worlds that satisfy α and are most normal among worlds satisfying α also satisfy β. (KLM, 169).

What corresponds to the "most preferred models" of a human reasoner? Arguably, they correspond to the representations that most accord with her prior beliefs and background knowledge about the world at a given time t, i.e., on the basis of the information available to her at t. As new information comes in, the agent may be led to revise or update her beliefs about the world; indeed, the preferential framework is inherently dynamic, and thus lends itself well to the representation of the development of an agent's cognitive states through time in function of the flow of information (van Benthem and Liu, 2004).

However, even though there is an avowed initial epistemic motivation at the heart of preferential logic, ultimately this is a framework developed within the artificial intelligence community, where goals of computational tractability and implementation remain central (alongside other desiderata and considerations). As such, the formal apparatus is accountable towards two potentially conflicting

12 The standard example goes: if A is "Tweety is a bird", the preferred models of this premise also validate the conclusion "Tweety can fly", but if C is "Tweety is an ostrich", then the class of preferred models of A and the class of preferred models of $\{A, C\}$ are disjoint.

sets of desiderata: epistemic plausibility vs. tractability (technical simplicity). One idealization related to tractability is the Limit Assumption, which guarantees that we are not dealing with a non-terminating chain of preferred models. This assumption ensures that a preferential logic enjoys the property of Cautious Monotony (if a set of premises imply A and the same set of premises imply B, then this set of premises plus A will also imply B), which is widely seen as a desirable metalogical property (Koons, 2013, sections 4.2 and 5.2). Now, this seems to be a desideratum mostly related to the *tractability* of the formal framework and its status as a logic, not to its *epistemic plausibility.*

We now explore how the framework can be applied to account for the empirical data on human reasoning presented above; we will see that some of its features do seem to limit the applicability of the preferential framework to the empirical data in significant ways.

As noted above, there are two patterns of discrepancy in participants' responses with respect to the deductive canons, both connected to the (un)believability of the conclusion: *overgeneration* – they draw inferences to "conclusions" that do not follow deductively from the premises but which are highly believable; and *undergeneration* – they refuse to draw inferences to conclusions that do follow deductively from the premises but which are highly unbelievable. Preferential logic fares well with the overgeneration phenomenon, but fails to explain the undergeneration phenomenon. This follows immediately from the observation that preferential logic is *supraclassical* (with respect to the original classical, monotonic logic L): for every A and B, if $A \Rightarrow B$, then $A \Rightarrow_{\prec} B$, given that the preferred models of A are also models of A *tout court*. So obviously, there are consequences that are preferentially but not classically valid, but not vice versa.

Before we move on to the empirical data, it may be worth pointing out again that the present analysis is above all concerned with the *descriptive* level of how human agents in fact reason, leaving aside the (thorny) *normative* question of how humans *ought* to reason.[13] The question is thorny because the tendency towards doxastic conservativeness and the "jumping to conclusions" phenomenon – in short, taking into account background knowledge when reasoning – are for the most part advantageous for human reasoners, but on certain occasions they also seem to lead to suboptimal results. But for reasons of space, we cannot discuss the normative issue any further here.

13 Harman (1986) has offered compelling arguments on why classical deductive logic is not a suitable normative system for human reasoning, but this does not necessarily mean that any of the non-classical alternatives available (either supraor subclassical) fares substantially better as a normative (as opposed to descriptive) model for reasoning.

Turning to the experimental results now, let us start with the easier case, overgeneration. Recall the following argument:

ψ: Some of the women are not beautiful.
ϕ: All of the beautiful people are actresses.
χ: Thus, some of the women are not actresses.

Following the semantic definition of preferential consequence, the question is: does χ hold in the preferred models of $\{\psi, \phi\}$? We know that it is not true in *all* models of $\{\psi, \phi\}$, and this is why χ is not a deductive consequence of ψ and ϕ. But among the models of $\{\psi, \phi\}$, which ones are considered "more normal" by the agent: the ones where χ holds or the ones where χ does not hold? Clearly the former are considered "more normal", even if $\{\psi, \phi\}$ is compatible both with χ and with not-χ. And so, in all preferred models of $\{\psi, \phi\}$, χ is also true, and thus $\{\psi, \phi\} \Rightarrow_{\prec} \chi$. Interestingly, this account resembles closely the "one-model" account of belief bias presented in Klauer et al. (2000), which is quite robustly supported by the data. So at this point preferential logic seems to fare well at the test of empirical adequacy posed by the belief bias data.

A similar argument holds for the other case of overgeneration discussed above:

All living things need water.
Roses need water.
Thus, roses are living things.

This argument also satisfies the definition of preferential consequence if is granted that in the agent's most preferred models of the premises, "Roses are living things" is satisfied.

By contrast, the preferential approach does not have much to say if the agent does not have any background knowledge or prior beliefs about the content of premises or conclusion, as in the hudon/wampets case mentioned above: in her preferred models, the conclusion is indeterminate. So she cannot resort to preferential reasoning to judge the validity of this argument. Presumably, this means that some other reasoning strategy must be called upon, which explains the discrepancy in results between the roses case and the wampets case (despite the similarity between the two cases from the point of view of classical/traditional logic).

Let us now turn to undergeneration. To account for the phenomenon of undergeneration in the same manner, what would be required is a logic that does *not* license some of the inferences licensed by classical logic – in other words, a *subclassical* logic. Because they are supraclassical, preferential logics are not able to account for why participants refuse to draw a deductive conclusion when it is unbelievable. However, the subclassical logics currently available, such as intuitionistic

and relevant logics, are strictly subclassical and thus do not license any inferences not licensed by classical logic. For this reason, they cannot handle overgeneration. Moreover, what seems to be going on in the undergeneration cases discussed here is not the kind of phenomenon that motivates these subclassical logics (e.g., rejection of *ex falso quodlibet* or of excluded middle). So we will need to look elsewhere. What could then explain the participants' refusal to draw the conclusion "Some of the actresses are not women" in the example above?

We submit that this tendency is related to the process of *revising* (or not) the agent's belief set with the given premises. In particular, the premise "All of the women are beautiful" is presumably not satisfied in the agent's belief set (i.e., the agent does not believe that all women are beautiful), which thus would require a revision process. Now, one peculiar feature of deductive reasoning is that reasoners are expected to reason with premises regardless of their belief or knowledge of them. This is, however, a somewhat artificial cognitive task for most (untrained) reasoners (Dutilh Novaes, 2013), and the artificiality component is enhanced in an experimental setting. Indeed, there is compelling empirical evidence to the effect that plausibility monitoring is a routine component of language comprehension (Isberner and Kern-Isberner, 2016), which underscores the "artificiality" of deductive reasoning as classically construed.

If the agent did indeed perform the revision of her belief set with the information "All of the women are beautiful" while retaining the belief that some actresses are not beautiful, then in her revised belief set, the counterintuitive conclusion – "Some actresses are not women" – would hold. But this process would require too much revision of the original belief set (including the very definition of "actresses"), and this seems to be what many participants in fact do not execute, despite being told to do so by the experimental instructions. (The idea that new information does not necessarily bring about a revision, e.g., if it is inconsistent with one's background knowledge, is precisely the rationale for the notions of *semi-revision* and *screened revision* introduced in the belief revision literature (Makinson, 1997), which will be discussed in the next section.)

It might be objected that, since "Some actresses are not women" is *definitionally* false, the agent does well to reject the conclusion, simply because this is the best way to avoid an inconsistent set of beliefs/commitments. But notice that undergeneration occurs also in cases where the conclusion is not definitionally false but simply highly unbelievable.[14] Indeed, in the Evans et al. (1983) study mentioned, deductively valid arguments with (merely factually) unbelievable conclusions such as "Therefore, some vitamin tablets are not nutritional" were also

14 Admittedly, in the Oakhill and Johnson-Laird (1985) study, the undergeneration effect was pronounced especially in cases of conclusions that were definitionally false.

deemed invalid in almost half of the cases. Moreover, a non-negligible 38% of participants did choose "Some actresses are not women" as the correct answer, thus apparently following the experiment's instructions closely.

This observation confirms yet again the general idea that humans have a tendency towards doxastic conservativeness and thus resist engaging in such revisions. This is the case specially when the revisions required are quite substantive and would thus entail significant modifications in their doxastic states – certainly when there is no real motivation to undertake such a revision, as in the case of an "artificial" reasoning experiment. But in fact, the tendency towards ignoring counterintuitive evidence, i.e., evidence which would force a thorough revision of one's prior beliefs, is also observed in more "natural" situations of reasoning, as the literature on confirmation bias illustrates (Nickerson, 1998).

In sum, it seems that the phenomenon of overgeneration of "conclusions" can be explained in terms of preferential reasoning, whereas the phenomenon of undergeneration of conclusions requires that attention be paid to the stage of revising one's belief set with incoming information (the premises). In cases where the incoming information is implausible (such as "all women are beautiful"), the agent may fail to perform the revision (either consciously or not), and this leads to the rejection of a counterintuitive conclusion which would however hold in her revised state of beliefs, had the agent in fact revised her belief set with the implausible premise. But as it stands, the preferential framework does not seem to offer the resources to tackle the issue of preferred models not being modified with the arrival of new information. (In fact, logic-based frameworks in general do not problematize what happens at the stage of receiving the premises.)

4. Undergeneration and screened revision

In the previous section we have argued that undergeneration cannot be straightforwardly explained by preferential logic because it is supraclassical. However, although it is indeed the case that no logical framework that satisfies supraclassicality can explain undergeneration, the problem is in fact deeper. We conjectured that in the case of undergeneration, the agent does not in fact incorporate the offered premises. So, the only way to restrict the "prioritized" status of incoming information is by giving up the reflexivity axiom, which says that A implies A, for any A. However, giving up reflexivity comes at a high price. Reflexivity is considered "a rather minimal requirement on a relation of logical consequence. It is hard to imagine in what sense a relation that fails to satisfy reflexivity, can still be considered a *consequence* relation" (Antonelli, 2012, section 1). Although

there might well be plausible consequence relations that are not reflexive,[15] all well-known non-monotonic consequence relations satisfy reflexivity. For this reason, in what follows we examine a different framework, namely the framework of belief revision, to investigate how it might deal with the undergeneration cases.

As is well known, theories of belief revision take a collection of beliefs, represented by sentences, as the belief state of an agent. Subsequently, operators are defined to perform the addition and removal of other beliefs. The main operators are revision and contraction. Since these operations might conflict with other beliefs of the agents, it may be necessary to perform changes on other beliefs as well. There are several frameworks that describe "how to revise a knowledge system in the light of new information that is inconsistent with what is already in the system" (Gärdenfors and Makinson, 1988, p. 83). In the original belief revision model (the AGM framework; Alchourrón et al., 1985), revision is performed using an operation called partial meet revision. The important thing for our purposes is that an operator is a partial meet revision operator if and only if it satisfies the following six postulates:

Closure $K * p = Cn(K * p)$
Success $p \in K * p$
Inclusion $K * p \subseteq K + p$
Vacuity If $\neg p \notin K$, then $K * p = K + p$
Consistency $K * p$ is consistent if p is consistent
Extensionality If $(p \leftrightarrow q) \in Cn(\emptyset)$, then $K * p = K * q$

K is the agent's belief set and is supposed to be closed under logical consequence, p is the input sentence and Cn is a supraclassical consequence relation. For our purposes, it is crucial to notice that there exists a translation between these postulates and the KLM system (Makinson and Gärdenfors, 1991). So, in the AGM framework, the belief revision operator can be considered to be functionally equivalent to the preferential consequence relation.

As has become clear in the previous section, any straightforward, supraclassical non-monotonic consequence relation will be at odds with the empirical data on belief bias in virtue of the undergeneration side. Now, if the AGM postulates are equivalent to the KLM system, then they will presumably not be able to account for undergeneration either. In particular, there is a problem with the second postulate, *Success*, which states that the input sentence is always an element of the resulting belief set (Makinson and Gärdenfors, 1991, p. 192). Importantly, the

15 In fact, the very first logical system ever invented, Aristotle's syllogistic, is irreflexive (Duncombe, 2014). But the consensus among logicians and philosophers still seems to be that reflexivity is a *sine qua non* condition for something to count as a legitimate consequence relation.

Success postulate has been considered empirically implausible in its own right, which makes the rejection of this postulate non-adhoc. *Success* is implausible, "because in many cases it is not reasonable to give priority to information just because it is new" (Simari and Falappa, 2004, p. 1342). Theories of belief revision that reject *Success* are generally called "non-prioritized".[16]

There are several ways to perform a non-prioritized revision operation. In this article we focus on a family of revision theories where the input sentence is first evaluated, and a "regular" prioritized belief revision is performed if and only if the sentence is considered acceptable (Hansson, 1999, pp. 413–414). A form of non-prioritized belief revision that works in this way is Makinson's *screened revision* (Makinson, 1997). This operation assumes a set A of core beliefs that are immune to revision. An input sentence is added to the core belief set of the agent if and only if it is consistent with A. If it is consistent, a revision operator can be used to revise the belief set. Otherwise, the input sentence is rejected.

This is a natural way to let background knowledge play an important role in determining whether a revision should be performed, but as Makinson recognizes, this "simple" version of screened revision, with a set of beliefs that are carved in stone, represents a rather dogmatic mentality. A somewhat more intuitive and flexible notion of belief revision is what Makinson calls *relationally screened revision*.[17] The idea behind this kind of revision is that an input is only accepted if it is consistent with the sentences from the original belief set that are *a priori* more credible than the input sentence (Makinson, 1997, p. 18). The notion of *a priori* credibility can be defined in different ways, but one way to do so is by using the notion of entrenchment (Gärdenfors and Makinson, 1988). If we denote *a priori* credibility (or entrenchment) with the $<$ operator, then relationally screened revision can be defined as follows:

Relationally screened revision (Makinson, 1997, p. 18)

$$K \#_{<}\alpha = K *_{\{\beta\,:\,\alpha<\beta\}}\alpha, \text{ if } \alpha \text{ is consistent with } \{\beta : \alpha < \beta\} \cap K$$
$$= K \text{ otherwise}$$
where $K *_{\{\beta\,:\,\alpha<\beta\}}$ is a revision operator that protects $\{\beta : \alpha < \beta\}$

16 An anonymous referee suggests that *Success* is not implausible if one preprocesses incoming information. The idea is that the prioritized revision operation should only be applied to sentences that are deemed believable in this first stage. We think this is correct, but non-prioritized belief revision is precisely a way to incorporate this "preprocessing" stage into the system.

17 In the first decade of this century, many alternative non-prioritized revision operators have been developed (see, for example, Fermé and Hansson, 2011, pp. 310–311, for a classification and Fermé and Rott, 2004, for an example of an alternative belief revision model that operates on the basis of epistemic entrenchment). Our focus on Makinson's relatively simple model should not be taken to imply that it is the only belief revision model that is empirically adequate with respect to the belief bias data.

This definition says that if the new information α is inconsistent with the beliefs from K that are *a priori* more credible than α, then no change occurs in K. If, on the other hand, this inconsistency is not present, it might still be the case that α is inconsistent with the whole body of K. In this case an operation is performed that protects the beliefs from K that are *a priori* more credible than α. So in either case, the beliefs from K that are *a priori* more credible than α remain present in the target belief set. Notice that the notion of entrenchment thus defined is very much in the spirit of the informal notion of doxastic conservativeness, which we claim is the overall phenomenon behind the belief-bias results.

In the remainder of this section, we show that relationally screened revision can explain both overgeneration and undergeneration. Hence, it is more empirically adequate with respect to the experimental data on belief bias than preferential logic.[18] Recall that in the typical example of overgeneration, we have the following argument:

ψ: Some of the women are not beautiful.
ϕ: All of the beautiful people are actresses.
χ: Some of the women are not actresses.

For most agents, ψ and χ are already part of their belief sets. Therefore, if the agent is presented with ψ, no changes will occur. If the agent is presented with ϕ, probably a new piece of information, it is necessary to determine whether ϕ is consistent with the beliefs that are *a priori* more credible than ϕ ($\{\beta : \phi < \beta\} \cap K$). If ϕ is not consistent with $\{\beta : \phi < \beta\} \cap K$, then ϕ is rejected. Otherwise, a revision is performed that protects all elements from $\{\beta : \phi < \beta\}$. Since $\chi \in K$ and $\phi < \chi$, χ is also an element of the resulting belief set.

Concerning undergeneration, the case that was not predicted by preferential logic, we are considering the following valid argument:

ψ: Some of the actresses are not beautiful.
ϕ: All of the women are beautiful.
χ: Some of the actresses are not women.

18 An anonymous referee suggests that it is possible to define a preferential consequence relation $\Rightarrow_{\prec}$ in terms of the relationally screened revision operator by using an adapted form of the Ramsey test: $A \Rightarrow_{\prec} B$ iff $A\#_{\prec}B$. So it might seem that, if there is a relationally screened revision operator which satisfies undergeneration, there must also be a preferential consequence relation that is as successful in the undergeneration cases. However, as was pointed out in the beginning of this section, this relation would not satisfy the reflexivity axiom, which for most people still is a non-negotiable feature for a consequence relation.

Since the agent probably has the more plausible "it is *not* the case that all of the women are beautiful" in his belief set, the expectation is that ϕ will be rejected. And if ϕ is not integrated into the agent's belief set, then χ will not be a consequence of the resulting (unmodified) belief set, and thus can be rejected by the agent.

It is interesting to notice that, in this particular experiment, most participants apparently went through the "screening" process leading to the non-incorporation of the premise provided, despite being told (explicitly or implicitly) to reason with the premises regardless of their plausibility. (We leave aside the issue of the extent to which this process is conscious or unconscious.) However, a non-negligible 38% of the participants *did* draw the highly counterintuitive conclusion, and thus presumably did perform the revision required to accommodate ϕ (even if only temporarily, and for the sake of the task at hand) into their belief sets. (More on individual variations in the next section.)

5. Observations

Initially, the main argument in favour of the preferential approach was that the key concept of "most preferred models" seems to correspond neatly to some robust psychological phenomena identified experimentally, and which have been described under different names: confirmation bias, belief bias, Stanovich's fundamental computational bias, Kahneman's WYSIATI, etc. Even if these are not exactly the same phenomena, they are all closely related, and all point in the direction of what we have described as a strong component of doxastic conservativeness in human cognition. However, because it offers no tools to problematize the very process of revising (or not!) an agent's belief state with incoming information, preferential logic offers at best a partial account of the phenomena. This is then reflected in its inability to account for the undergeneration part of the story: why some agents often refuse to draw conclusions that are counterintuitive.

And yet, undergeneration and overgeneration both arguably arise from the same tendency towards doxastic conservativeness, which leads human reasoners to "jump to conclusions" when they are plausible, and to reject conclusions which are implausible. This means that neither supraclassical nor subclassical logics can fully match the empirical data in question: supraclassical logics such as preferential logic are at odds with the undergeneration phenomenon, while subclassical logics are at odds with the overgeneration phenomenon. Above, we offered a brief sketch of how screened revision is better equipped to deal with both phenomena in a uniform way.

In fact, an alternative account of the belief bias data discussed here is the following: what participants are in fact doing in the experiments is not reasoning at all, at least not in the sense of drawing conclusions from premises. Instead, they seem to be engaging in what could be described as *belief management*: what happens when new information comes in? Which beliefs do I still hold on to? Reasoning experiments, and indeed deductive reasoning in general, ask of participants to perform an "artificial" revision of their beliefs, i.e., to accept the premises for the sake of the argument. We propose that a thus far not sufficiently discussed aspect of these results is the extent to which participants in fact do or do not perform a revision in the case of unbelievable premises. Thus, the non-prioritized belief revision formal framework as presented here suggests that the effect of beliefs in the participants' performance, which is viewed as a "bias" under the assumption that participants should be performing deductive reasoning, is perhaps no bias at all if what they are actually doing is engaging in belief management in view of background knowledge.

Finally, it also bears emphasizing that, in the reasoning experiments discussed above (and in fact, in most reasoning experiments), there is considerable *individual variation* in responses. For example, in the undergeneration example discussed above where the syllogistic conclusion to be drawn was "Some of the actresses are not women", a significant group of 38% participants *did* draw this highly counterintuitive but deductively correct conclusion. (We hypothesize that these different responses are related to participants performing or not performing the revision required by the false premise. Whether an individual does or does not perform the revision will depend on her credibility ordering.) Indeed, any plausible account of human reason must also make room for the phenomenon of inter-personal variation.

6. Conclusion

In this article, we began by investigating how well the framework of preferential logic fared when compared to some very robust experimental data on human reasoning, in particular data pertaining to the phenomena described under the heading of confirmation/belief bias. The initial motivation was the observation that the concept of preferred models, which is the conceptual cornerstone of preferential logic, seems to capture an important feature of human cognition, namely that (contrary to the precepts of deductive reasoning) we only take into account situations that are minimally plausible (from the point of view of our prior beliefs and background knowledge) for the purposes of reasoning. We showed that preferential logic makes accurate predictions related to how human reasoners "jump to

conclusions" on the basis of limited available information, thus tentatively drawing conclusions that are not deductively warranted by the available information (overgeneration).

However, preferential logic was also found to make the wrong predictions with respect to the phenomenon of undergeneration, i.e., the fact that reasoners may refuse to draw deductively warranted conclusions if they clash with their prior beliefs. We suggested that undergeneration requires that attention be paid to what happens when new information arrives and the agent must decide (consciously or not) whether (or not) to perform a revision in her belief state with the incoming information. This aspect in turn is treated more successfully by the screened belief revision framework. The empirical adequacy of screened belief revision with respect to the belief bias experimental data suggests that what participants are engaging in in the experiments is something that could be described as belief management, rather than the act of drawing conclusions from premises. Interestingly, in cases where there are no beliefs to speak of related to premises and conclusion (say, in the wampet/hudon example), then belief management is not (cannot be!) what participants engage in, and presumably they may then engage in something closer to deductive reasoning (recall that the wampet/hudon case elicited 78% of correct responses).

We thus conclude that preferential logic does not pass the test of empirical adequacy posed by the belief bias data; in fact, any supraclassical logic will run into the same difficulties when confronted with undergeneration. In contrast, screened belief revision predicts a fair amount of both undergeneration and overgeneration. In both frameworks, prior belief and background knowledge play a crucial role, but in different ways: in preferential logic, an agent's existing belief set will determine her preference ordering for models, whereas in screened belief revision it becomes activated at the initial, screening stage. Screened revision makes use of core beliefs, which must be preserved (though in relationally screened revision there is not necessarily a set of core beliefs which will not change, no matter what), whereas in standard preferential reasoning, background knowledge can be more easily revised.

Moreover, the present investigation can be seen as a case study for the fruitfulness but also the limitations of studying empirical phenomena pertaining to human cognition from the point of view of formal frameworks. Formal frameworks can lead to the formulation of hypotheses and experiments, and to the (partial) explanation of the results. However, it may also happen that properties of the formal framework in question, such as supraclassicality and the reflexive consequence relation in preferential logic, do not reflect adequately the relevant empirical phenomena, as we have argued here.

References

Alchourrón, C. E., Gärdenfors, P. and Makinson, D. (1985) "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions." *Journal of Symbolic Logic*, 510–530.

Antonelli, G. A. (2012) "Non-monotonic Logic." In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, http://plato.stanford.edu/archives/win2012/entries/logic-nonmonotonic/.

Benferhat, S., Bonnefon, J. F. and da Silva Neves, R. (2005) "An Overview of Possibilistic Handling of Default Reasoning, with Experimental Studies." *Synthese* 146(1–2): 53–70.

Duncombe, M. (2014) "Irreflexivity and Aristotle's Syllogismos." *Philosophical Quarterly* 64(256): 434–452.

Dutilh Novaes, C. (2013) "A Dialogical Account of Deductive Reasoning as a Case Study for How Culture Shapes Cognition." *Journal of Cognition and Culture* 13(5): 459–482.

Evans, J. S. B. (1989) *Bias in Human Reasoning: Causes and Consequences*. Lawrence Erlbaum Associates.

Evans, J. S. B., Barston, J. L. and Pollard, P. (1983) "On the Conflict between Logic and Belief in Syllogistic Reasoning." *Memory & Cognition* 11(3): 295–306.

Fermé, E. and Hansson, S. O. (2011) "AGM 25 years." *Journal of Philosophical Logic* 40(2): 295–331.

Fermé, E. and Rott, H. (2004) "Revision by Comparison." *Artificial Intelligence* 157(1–2): 5 – 47.

Gärdenfors, P. and Makinson, D. (1988) "Revisions of Knowledge Systems Using Epistemic Entrenchment." In *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 83–95. Burlington, MA: Morgan Kaufmann.

Hansson, S. O. (1999) "A Survey of Non-Prioritized Belief Revision." *Erkenntnis* 50(2–3): 413–427.

Harman, G. (1986) *Change in View*. Cambridge, MA: MIT Press.

Isberner, M.-B. and Kern-Isberner, G. (2016) "A Formal Model of Plausibility Monitoring in Language Comprehension." In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, FLAIRS-29*, pp. 662–667. Palo Alto, CA: AAAI Press.

Kahneman, D. (2011) *Thinking, Fast and Slow*. London: Penguin.

Klauer, K. C., Musch, J. and Naumer, B. (2000) "On Belief Bias in Syllogistic Reasoning." *Psychological Review* 107(4): 852–884.

Klayman, J. (1995) "Varieties of Confirmation Bias." *Psychology of Learning and Motivation* 32: 385–418.

Koons, R. (2013) "Defeasible Reasoning." In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Spring 2013 edition.

Kraus, S., Lehmann, D. and Magidor, M. (1990) "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics." *Artificial Intelligence* 44(1): 167–207.

Makinson, D. (1997) "Screened Revision." *Theoria* 63(1–2): 14–23.

Makinson, D. and Gärdenfors, P. (1991) "Relations between the Logic of Theory Change and Nonmonotonic Logic." In A. Fuhrmann and M. Morreau (eds), *The Logic of Theory Change*, pp. 183–205. Dordrecht: Springer.

Nickerson, R. S. (1998) "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2(2): 175.

Oakhill, J. and Johnson-Laird, P. N. (1985) "The Effects of Belief on the Spontaneous Production of Syllogistic Conclusions." *Quarterly Journal of Experimental Psychology* 37 (4): 553–569.

Oaksford, M. and Chater, N. (1991) "Against Logicist Cognitive Science." *Mind & Language* 6(1): 1–38.

Pfeifer, N. and Kleiter, G. D. (2005) "Coherence and Nonmonotonicity in Human Reasoning." *Synthese* 146(1–2): 93–109.

Quine, W. V. (1955) "Posits and Reality." Reprinted in *The Ways of Paradox and Other Essays*, Vol. 2, pp. 246–254. Cambridge MA: Harvard University Press.

Sá, W. C., West, R. F. and Stanovich, K. E. (1999) "The Domain Specificity and Generality of Belief Bias: Searching for a Generalizable Critical Thinking Skill." *Journal of Educational Psychology* 91(3): 497.

Shoham, Y. (1987) Readings in Nonmonotonic Reasoning. chapter A Semantical Approach to Nonmonotonic Logics, pp. 227–250. San Francisco: Morgan Kaufmann.

Simari, G. I. and Falappa, M. A. (2004) "Non Prioritized Belief Revision with Ansprolog." In *VI Workshop de Investigadores en Ciencias de la Computación*, http://sedici.unlp.edu.ar/handle/10915/21236.

Stanovich, K. E. (2003) "The Fundamental Computational Biases of Human Cognition: Heuristics That (Sometimes) Impair Decision Making and Problem Solving." In J. E. Davidson and R. J. Sternberg (eds), The Psychology of Problem Solving, pp. 291–342. Cambridge: Cambridge University Press.

Stenning, K. and Van Lambalgen, M. (2008) *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Stenning, K. and Van Lambalgen, M. (2010) "The Logical Response to a Noisy World." In M. Oaksford and N. Chater (eds), *Cognition and Conditionals: Probability and Logic in Human Thinking*, pp. 85–102. Oxford: Oxford University Press.

Tversky, A. and Kahneman, D. (1974) "Judgment under Uncertainty: Heuristics and Biases." *Science* 185(4157): 1124–1131.

van Benthem, J. and Liu, F. (2004) "Diversity of Logical Agents in Games." *Philosophiae Scientiae* 8(2): 163–178.

Van Fraassen, B. C. (1980) *The Scientific Image*. Oxford: Oxford University Press.