# Belief Change as Propositional Update

Renée Elio

Francis Jeffry Pelletier

University of Alberta

Edmonton, Alberta

T6G 2H1

# Abstract

In this study, we examine the problem of belief revision, defined as deciding which of several initially-accepted sentences to disbelieve, when new information presents a logical inconsistency with the initial set. In the first three experiments, the initial sentence set included a conditional sentence, a non-conditional sentence, and an inferred conclusion drawn from the first two. The new information contradicted the inferred conclusion. Results indicated that the conditional sentences were more readily abandoned than non-conditional sentences, even when either choice would lead to a consistent belief state, and that this preference was more pronounced when problems used natural language cover stories rather than symbols. The pattern of belief revision choices differed depending on whether the contradicted conclusion from the initial belief set had been a modus ponens or modus tollens inference. Two additional experiments examined alternative model-theoretic definitions of minimal change to a belief state, using problems that contained multiple models of the initial belief state and of the new information that provided the contradiction. The results indicated that people did not follow any of four formal definitions of minimal change on these problems. The new information and the contradiction it offered was not, for example, used to select a particular model of the initial belief state as a way of reconciling the contradiction. The preferred revision was to retain only those initial sentences that had the same, unambiguous truth value within and across both the initial and new information sets. The study and results are presented in the context of certain logic-based formalizations of belief revision, syntactic and model-theoretic representations of belief states, and performance models of human deduction. Principles by which some types of sentences might be more "entrenched" than others in the face of contradiction are also discussed from the perspective of induction and theory revision.

# Belief Change as Propositional Update

Suppose you need to send an express courier package to a colleague who is away at a conference. You believe that whenever she is in New York City and the New York Rangers are playing a home game, she stays at the Westin Mid-Manhattan Hotel. You also believe that she is in New York City this weekend and that the Rangers are playing this weekend as well. You call up the Westin Mid-Manhattan Hotel and you find out that she isn't there. Something doesn't fit. What do you believe now? Well, assuming that you accept the hotel's word that she isn't there, there are various (logically consistent) ways to reconcile the contradiction between what you used to believe and this new information. First, you could believe that she is in New York City and that the Rangers are indeed playing, but disbelieve the conditional that says whenever both of these are true, then she stays at the Westin Mid-Manhattan Hotel. Alternatively, you could continue to believe the conditional, but decide that either she isn't in New York this weekend or that the Rangers aren't playing a home game (or possibly both). Which do you choose as your new set of beliefs?

Belief change—the process by which a rational agent makes the transition from one belief state to another—is an important component for most intelligent activity done by epistemic agents, both human and artificial. When such agents learn new things about the world, they sometimes come to recognize that new information extends or conflicts with their existing belief state. In the latter case, rational reasoners would identify which of the old and new beliefs clash to create the inconsistency, decide whether in fact to accept the new information, and, if that is the choice, to eliminate certain old beliefs in favor of the new information. Alternatively, new information may not create any inconsistency with old information at all. In this case, the reasoner can simply add the new information to the current set of beliefs, along with whatever additional consequences this might entail.

Although this is an intuitively attractive picture, the principles behind belief-state change are neither well-understood nor agreed-upon. Belief revision has been studied from a formal perspective in the artificial intelligence (AI) and philosophy literatures and from an empirical perspective in the psychology and management-science literatures. One of the practical motivations for AI's concern with belief revision, as portrayed in our opening scenario, is the development of knowledge bases as a kind of intelligent database: one enters information into the knowledge base and the knowledge base itself constructs and stores the consequences of this information—a process which is non-monotonic in nature (i.e., accepted consequences of previously-believed information may be abandoned). More generally, the current belief state of any artificial agent may be contradicted either when the world itself changes (an aspect of the so-called frame problem) or when an agent's knowledge about a static world simply increases. Katsuno and Mendelson (1991) distinguish between these two cases, calling the former belief <u>update</u> and latter belief <u>revision</u>. Although much of the AI belief revision work focuses on formalizing competence theories of update and revision, prescriptive principles for how artificial agents “should” resolve conflict in the belief revision case—where there is a need to contract the set of accepted propositions in order to resolve a recognized contradiction—are far from settled. From the perspective of human reasoning, we see an important interplay between issues of belief revision and deductive reasoning, particularly in terms of the kind of representational assumptions made about how a belief state should be modeled. But while human performance on classical deductive problems has been extensively studied, both Rips (1994, p. 299) and Harman (1986, p. 7) have noted the need for descriptive data and theories on how people resolve inconsistency when new information about a static world is presented. The studies we present in this article are concerned exactly with this issue.

We make two simplifications in our portrayal of belief revision and the paradigm we used to investigate it. The first concerns what we refer to as "beliefs." Here, beliefs are sentences that people are told to accept as true, in the context of resolving some (subsequent) contradiction arising from new information that is provided. Being told to accept something as true is not necessarily the same as believing it to be true. The contradictions we introduce in our paradigm are not probes into a person's pre-existing belief system (e.g., as in social cognition investigations of attitude change; see Petty, Priester, & Wegener, 1994) or of a person's hypotheses that are acquired over time via direct interactions with the world. The second simplification we make is treating beliefs as propositions that are believed either to be true or to be false (or, sometimes, that have a belief status of "uncertain"). This idealization characterizes the perspective of AI researchers who are interested in showing how classical deductive reasoning is related to belief revision. We will call this perspective "classical belief revision," to distinguish it from other frameworks, including one direction in formal studies of defeasible reasoning, that map statistical or probabilistic information about a proposition into a degrees of belief in that proposition (Kyberg, 1983, 1994; Bacchus, Grove, Halpern, and Koller, 1992; Pollock, 1990; Pearl, 1988). Both classical belief revision and defeasible reasoning are concerned with non-monotonicity and it is possible to view belief revision as driving defeasible reasoning or vice versa (Gärdenfors, 1990a; Makinson & Gärdenfors, 1991).

This alternative formalization of beliefs and belief change in terms of probabilistic or statistical information have analogies in certain empirical investigations as well. A primary concern in the management-science literature, for example, is to understand what factors influence a shift in the degree of belief in a particular proposition of interest. These factors include information framing (e.g., Ashton & Ashton, 1990; Shields, Solomon, & Waller, 1987) and agreement with prior beliefs and expectations (e.g., Koehler, 1993). Carlson and Dulany (1988) have proposed a model of belief revision about causal hypotheses from circumstantial evidence, in which the level of certainty in a causal hypothesis depends in part on the level of certainty the reasoner ascribes to circumstantial evidence supporting it. In Thagard's (1989) computer model of explanatory coherence, propositions have levels of activation that roughly correspond to acceptance levels; such a model has been applied to accounts of scientific reasoning and to belief revision as evidenced in protocols of subjects performing elementary physics (Ranney & Thagard, 1988).

Notwithstanding these alternative ways to conceptualize belief states, we believe that the issues investigated under our simplifications are relevant to these other perspectives. Belief revision as a deliberate act by an agent must be driven by something, and that driving force must include the detection of a conflict (defined logically or otherwise) within the belief state. The problem of explicitly "expunging" or contracting of beliefs, after having noticed a conflict, has been acknowledged within some degree-of-belief frameworks (e.g., Kyberg, 1983; 1994)**.** As soon as one attempts to define notions like "acceptance" or "full commitment to" within a degrees-of-belief framework, for the purpose of making a decision or taking an action, then new information can introduce conflict with existing accepted information. Hence, the issue still remains as to which prior belief or assumption an agent continues to believe (or to increase the degree of belief in) and which the agent decides to abandon (or decrease the degree of belief in). [1]

Belief revision has also been studied as something that does not occur when it “should.” That is, there is considerable evidence indicating that people are in general very reluctant to change their current belief sets in the face of evidence that indicates those beliefs are unjustified; and that they are much more likely to reject, ignore, or reinterpret the new information which conflicts with their current beliefs rather than attempt to add it to their beliefs and make the necessary adjustments (Edwards, 1968; Einhorn & Hogarth, 1978; Ross & Lepper, 1980; Hoenkamp, 1988; Lepper, Ross, and Lau, 1986). Although it is true that there are occasions in which people fail to revise their beliefs or refuse to accept new information, and there are theories offered as accounts of that reluctance, our starting point in these investigations assumes that any inertia against changing a belief set has been overcome.

Given our simplifications for the representation of belief states, the specific issue that concerns us can be easily stated. It is the question of which belief(s) out of some initial set is (are) abandoned when new, contradictory information must be integrated. The matters we consider in this study relate to certain formal notions that have been central to (what we have called) the classical AI belief revision perspective. These notions are epistemic entrenchment (whether some forms or types of information are less readily abandoned to resolve contradiction) and minimal change. It is not possible to consider these ideas without considering the fundamental choice that theories make in modeling a belief state either as a set of formulae or as a set of models. The implications of choosing one framework or another are crucial to operationalizing ideas like epistemic entrenchment. We review these two alternative positions on modeling belief states, and their relation to theories of human deduction, in the next section.

## On Modeling Belief States and Deduction

### Classical Models of Belief Revision

Alchourrón, Gärdenfors, & Makinson (1985; henceforth, "AGM") proposed a set of "rationality postulates" as a competence specification of what rational belief change should be. Many of these ideas are intrinsically important to thinking about human belief revision as we are studying it here, so we borrow some key distinctions from that literature in setting the stage for our studies.

There are two predominant camps in how belief states are modeled within what we earlier defined as the classical belief revision community: "syntactic-based theories" v. "model-based theories". The majority of the work in either of these camps follow the idealizations we outlined above: that beliefs are propositional in nature, that the status of a belief is "believed true", "believed false" or "uncertain", and that logical inconsistency is to be avoided within the agent's chosen belief state.

The difference between the syntactic and model approaches can be seen by example. Consider what might be in a belief state when an agent told: All of Kim's cars are made in the US; This (some particular) car is made in Germany. The syntax-based theories take the position that what is stored in the agent's belief state are the two formulas mentioned (plus whatever background information the agent already had…also stored as a set of formulas). Since beliefs are just formulas, doing a logical inference amounts to performing some further mental activity on these formulas. This further activity would generate a different belief state from the initial one. And so there is no guarantee that the agent will perform any logical inferencing to generate new beliefs. For instance, there is no guarantee that this agent will use the background information it may have that Germany is a different country than the US and that cars made in the one are not made in the other to infer that this car is not made in the US. Even if it does perform this inference, there is no guarantee that it will make the further inference that the car is not owned by Kim. In this conception, two beliefs are different when and only when they are expressed by two syntactically distinct formulas.

In contrast to this, the model-based theories identify a belief state with a model—an interpretation of the world which would make a group of beliefs be true. In the above example of Kim's cars, a model-based theory would identify the agent's belief state with those models of the world in which all of Kim's cars are made in the US and where furthermore some particular car is made in Germany. Assuming the agent's background beliefs include that Germany is a different country than the US and that cars made in the one are not made in the other, the set of background models that can accommodate such situations in the world are merged with those describing the two stated beliefs and the output is a model (or models) in which Kim's cars are made in the US, and this car is made in Germany, and hence this car is not made in the US, and hence this car is not owned by Kim. All this sort of "inferencing" is done already in the very description of the belief state. The fact that the belief state is a model of the world described by the sentences guarantees that all logical consequences of these sentences will be represented, for otherwise it couldn't be a model of those sentences.

One common way of putting the difference is to say that the syntax-based approach is committed only to explicit beliefs as defining a belief state, whereas a model-based approach is committed to defining a belief state in terms not only of explicit beliefs but also of the implicit beliefs that are entailed by the explicit ones. Both approaches involve a certain amount of theoretical idealization. Under the model-based view, the very definition of an agent’s belief state already embodies finding the models that perfectly suit it, and this in effect means that all the logical conclusions of any explicit beliefs are included. Within the syntactic framework, there is an assumption that only "obvious" or "minimal" conclusions are drawn, but how these are recognized as such goes unspecified. Secondly, it is not clear how syntactic-based theories detect arbitrary logical contradictions beyond ones that can be immediately spotted by a syntactic pattern-match, such as "p and ~p", since beliefs are represented as strings of symbols and not models of the world being described. [2]

A third conception of belief states—which could be seen as an intermediate stance between the syntactic and the model-based approaches— might be called a “theory-based” theory of beliefs. Here a belief state is identified with a theory, which is taken to be a set of sentences, as the syntactic-based theories hold. However, this set is the infinite set of all the logical consequences of the explicit beliefs. [4] This is the approach advocated in the original work done by AGM (1985). It too is obviously an idealization, for taken to the extreme, it would require a person's mind (or an agent's memory) to be infinite in order to hold a belief. Although theory-based theories are like syntax-based theories in containing formulas (and unlike model-based theories in this regard), they differ from syntax-based theories in obeying a principle called “The Irrelevance of Syntax”: if two formulas are logically equivalent, then adding one of them to a belief state will yield the same result as adding the other, since the set of their logical consequences is the same. This principle is obeyed by both the theory-based and the model-based theories, and has been vigorously defended (AGM, 1985; Dalal, 1986; Yates 1990; Katsuno & Mendelson,

1991) on the grounds that all that is relevant to belief change is how the world is, or would be, if the beliefs were true.

Many of the concepts and distinctions mentioned above as characterizing classical AI belief revision also apply to other belief revision frameworks. Computational frameworks in which propositions are represented as nodes in some kind of belief network (e.g., Pearl, 1988; Thagard, 1989) are syntactic, insofar as any semantic contradiction between two nodes must be reflected in the names of the links chosen to join the nodes in the network. Methods proposed by Halpern (1990) and Bacchus et al. (1992) for deriving degrees of belief from statistical information are model-based approaches: the degree of belief in a sentence stems from the probability of the set of worlds in which the sentence is true. Kyberg's theory of rational belief (1983;1994), in which levels of acceptance are also derived from probabilities, falls into what we have called the "theory theory" category. He models beliefs as a set of (first-order logic) sentences but then requires the set to obey the irrelevance of syntax principle: if two sentences have the same truth value in belief set, then their probabilities are also equivalent within that set. So we see that, although the classical belief revision approach comes from a milieu were performance criteria are not explicitly considered, the sorts of distinctions made within these classical belief revision frameworks can elucidate the representational assumptions of other approaches as well.

## Performance Theories of Human Deduction

Harman (1986) has argued that the principles guiding belief revision are not the rules of deductive logic. Certainly, any principles that can dictate which of several different belief-state changes to select are outside the scope of deductive inference rules. Any characterization of belief revision must first make some commitment to how a belief state is represented; as the formal theories we outlined above illustrate, making (or not making) inferences is crucial to how the belief revision process is to be conceptualized.

Certainly, the ability to recognize inconsistency is a necessary step towards deliberate belief revision, and that step may involve some aspects of what has been studied and modeled in the laboratory as deductive reasoning. Hence, it seems that theories about how people draw inferences from propositional knowledge will be crucially related to the transition from one belief state to another, if only because those inferences may define the content of the belief states themselves.

Generally speaking, the theories of human deductive reasoning have split along a dimension that is similar to, but not identical with, the syntactic v. model-theoretic distinction in AI. On the one hand, mental-model theories of the type proposed by Johnson-Laird and colleagues (Johnson-Laird, Byrne, & Schaeken, 1992; Johnson-Laird & Byrne, 1991) hold that a person reasons from particular semantic interpretations (models) of sentences such as $\underline{p \rightarrow q}$ and *either p or q*.[3] In this framework, a reasoner identifies or validates a particular conclusion by manipulating and comparing these models. On the other hand, proof-theoretic approaches (Braine & O'Brian, 1991; Rips, 1983; 1994) propose that people possess general inference rules and follow a kind of natural deduction strategy to derive conclusions from a set of premises. Like the different kinds of belief revision theories in AI, these different psychological accounts of human deduction offer distinct representational assumptions about the constituent parts that are said to define a belief state. But unlike the AI belief revision theories, psychological theories must make a commitment to a plausible process account of how a person generates and operates upon these different representations.

Neither mental-model nor proof-theoretic accounts of deduction were initially developed for belief revision as we have portrayed it here; nor have there been, as yet, extensions designed to accommodate aspects of this phenomenon. However, we consider some of their basic assumptions in the discussion of our tasks and results, and so here we briefly summarize the mental-models framework proposed by Johnson-Laird and colleagues and the proof-theoretic model proposed by Rips (1994).

If we apply a state-space abstraction to mental models frameworks and to proof-theoretic frameworks, the main distinction between proof theoretic and model-based theories of human deduction can be summarized as differences in what defines a state and what constitutes the operators that make transitions from one state to another. In a proof-theoretic system like the one proposed by Rips (1994), a state is a partial proof and the operators are a set of inference rules (a subset of the classical logic inference rules). These operators extend a proof (and hence move the system from one state to the next) by following a natural deduction-like strategy, with heuristics that order their application within this general control strategy. The goal can be viewed as a state (or a path to a state) which includes a given conclusion as an outcome of a proof (hence validating it) or includes a statement not already specified in the problem's premises (drawing a new conclusion). In the mental-models theory, a state contains one or more interpretations of the sentence set, i.e., tokens with specific truth values that correspond to some situation in the world. Operators retrieve models of sentences and move the system to new states that constitute candidate models of the world. More specifically, the mental models framework assumes there are particular models that are initially associated with particular sentence forms (conditionals, disjuncts, and so forth), with other models of these forms sometimes held in abeyance until there is a need to consider them. A conclusion is any truth condition that is not explicitly stated in the sentence set, but which must hold given a consistent interpretation of the sentence set. Hence, the goal state can be seen as one in which such a truth condition is identified. Thus, the proof-theoretic theory of human deduction can be seen as a search for alternative inference rules to apply to a sentence set in order to extend a proof, whereas the mental models theory can be seen as a search for alternative interpretations of a sentence set, from which a novel truth condition can be identified or validated.

It is important to be clear not only about the similarities but also about the differences between the classical, competence belief-revision theories and the psychological performance theories of human deduction. What the mental-models theory shares with the formal model-based belief revision theories is the essential idea that the states being operated upon are models. These models capture the meaning of the connectives as a function of the possible truth values for individual atomic parts that the connectives combine. However, there are three key differences between these two types of model theories. First, although the irrelevance-of-syntax principle is a distinguishing feature of formal models of belief revision in AI, it does not distinguish between mental-models and proof-theoretic models of human deductive reasoning, which offer alternative accounts of the pervasive finding that syntactic form *does* influence how people reason about problems that are otherwise logically equivalent. Second, in the mental-models theory, the models of $p \rightarrow q$ are generated in a serial, as-needed basis, depending on whether a conclusion is revealed or validated by the initial interpretation (and it is the order in which such models are generated that plays in the mental-model's account of the effect of syntactic form on deductive reasoning). The AI model-based belief revision frameworks do not make any such process assumptions, except in their idealization that all models are available as the belief state. Third, the mental-models framework may be considered closer to what we have called the "theory theory" classical belief-revision framework, than to the pure model framework, because separate models of each sentence are produced and operated upon. What a psychological proof-theoretic framework of deduction shares with its formal AI syntactic-based counterparts is a commitment to apply deductively sound inference rules to sentences. But unlike the syntactic-based competence theories of belief revision, psychological proof-theoretic models of deduction do not presume that a person has a representation of every deductive rule of inference and they may presume there is some heuristic ordering of the available rules; these differences are relevant to how a proof-theoretic perspective models the relative difficulties that people have with certain forms of deductive problems. Further, some of the undesirable aspects of syntactic competence models, such as uncontrolled deductive closure, are avoided in proof-theoretic performance models (e.g., Rips, 1994) by explicitly positing that the reasoner's current goals and subgoals directs and controls the application of inference rules.

## Minimal Change and Epistemic Entrenchment

A basic assumption behind most AI theories of belief revision (e.g., the AGM postulates) and some philosophical accounts (e.g., Harman, 1986) is that an agent should maintain as much as possible of the earlier belief state while nonetheless accommodating the new information. But it is not completely clear what such a minimal change is. First, there is the problem in defining a metric for computing amounts of change. Often, this relies on counting the number of propositions whose truth value would change in one kind of revision versus another. The revision that leaves the belief state “closest” to the original one is to be preferred. But note that how such a definition of closeness works depends on whether one takes a syntactic or model-based approach.

As an example of the differences that can evolve, consider our earlier story about your New York-visiting colleague. Let $n$ stand for she-is-in-New-York, $r$ stand for Rangers-are-playing, and $w$ stand for she-stays-at-the-Westin. Symbolically, your initial beliefs were $[n \& r \rightarrow w, n, r]$, from which you deduced $w$. But then you found out $\sim w$. In a model-based approach, the unique model describing your initial belief set (unique, at least, if we attend only to $n$, $r$, and $w$) is: $n$ is true, $r$ is true, $w$ is true. Then you discover that the model is incorrect because $w$ is false. The minimal change you could make is merely to alter $w$'s truth value, and so your resulting belief states is: $n$ is true, $r$ is true, $w$ is false. In a syntax-based approach, you would instead keep track of the ways that as many as possible of the three initial sentences remain true when you add $\sim w$ to them. There are three such ways: $S_1= [n \& r \rightarrow w, n, \sim r, \sim w]$, $S_2=[n \& r \rightarrow w, \sim n, r, \sim w]$, $S_3=[\sim(n \& r \rightarrow w), n, r, \sim w]$.[5] Now consider what common sentences follow from each of $S_1$, $S_2$, and $S_3$, and the answer is that the consequences of $[\sim w, n \vee r]$ will describe them.

Note that this is different from the version given as a model-theoretic solution. In the syntactic case, only one of n and r need remain true, whereas in the model-based belief revision version, both need to remain true.

The notion of "epistemic entrenchment" in the belief revision literature (Gärdenfors, 1984, 1988; Gärdenfors & Makinson, 1988; Nebel, 1991; Willard & Yuan, 1990) has been introduced as a way to impose a preference ordering on the possible changes. Formally, epistemic entrenchment is a total pre-ordering relation on all the sentences of the language, and this ordering obeys certain postulates within the AGM framework. Less formally, epistemic entrenchment can be viewed as deeming some sentences as “more useful” and hence more entrenched against possible abandonment than other sentences; and in cases where there are multiple ways of minimizing a change to a new belief state, these priority schemes will dictate which way is chosen. Now, the general issue of whether some types of knowledge (e.g., sensory observations v. reasoned conclusions) should be a priori more epistemically privileged than other types of knowledge has occupied much of philosophy throughout its history. One particular, more modest, contrast is between what might be called statements about data v. statements about higher-order regularities. From one perspective, it can seem that conditional statements should enjoy a greater entrenchment in the face of conflicting evidence because of they express either semantic constraints about the world or express an important predictive regularity that might be the result of some long-standing and reliable inductive process. As an example of this sort of perspective, one can point to scientific theorizing that is based on statistical analysis, were one rejects "outlying" data as unimportant, if other regularities characterize most of the remaining data. In doing so, we give priority to the regularity over (some of) the data. Certain approaches to database consistency (e.g., Elmasri & Navathe, 1994, pp. 143-151) and some syntactic theories of belief revision (e.g., Willard & Yuan, 1990; Foo & Rao, 1988) advocate the entrenchment of the conditional form $p \rightarrow q$ over non-conditional forms. For database consistency, a relation like $p \to q$ can be said to represent a semantic integrity constraint, as in "*If x is y's manager, then x's salary is higher than y's salary*." For classical belief revision theories, the intuition driving the idea of entrenching $p \to q$ over other types of sentences is not because material implication *per se* is important, but because "lawlike relations" are often expressed in sentences of this form. For example, Foo and Rao (1988) assign the highest epistemic entrenchment to physical laws, which may be especially effective in reasoning about how a dynamic world can change (e.g., the belief *update*, rather than *revision*, problem).

But there is another perspective that would propose exactly the opposite intuitions about entrenchment: what should have priority are observations, data, or direct evidence. These are the types of statements which are fundamental and about which we can be most certain. Any kind of semantic regularities expressed in conditional form are merely hypotheses or data-summarizing statements that should be abandoned (or at least suspected) when inferences predicted from them are not upheld by direct evidence. This sentiment for data priority seems entirely plausible in the context of hypothesis evaluation (e.g., Thagard, 1989) as it did to some involved in the "logical construction of the world" (e.g., Russell, 1918; Wittgenstein, 1922).

In sum, we note that these alternative intuitions about entrenching conditionals v. non-conditionals are more or less readily accommodated, depending on the representation of belief states. It is easy to use the form of a sentence as a trigger for entrenchment principles, if one has a syntactic stance; but if a reasoner works with models of the world, then this sort of entrenchment is not as easily supported (unless sentences are individually modeled and knowledge of "form" is somehow retained). By first understanding the principles that actually guide belief revision in people, we are in a better position to formulate what kinds of representations would enable those principles to operate in a cognitive system.

## Overview of Experiments

So far, we have touched upon a number of broad theoretical issues that bear on belief revision, at least when this is characterized as a deliberate decision to remove some proposition(s) that had been accepted as true, in order to resolve a contradiction noted in the belief set. Although our longer-term interest is to better understand what plausible principles might define epistemic entrenchment, our immediate interest in the present studies was first to acquire some baseline data on what belief revision choices people make in relatively content-free tasks and to tie these results to models of deduction. To do this, we consider the simple task of choosing to abandon a conditional proposition v. a non-conditional proposition (what we will also call a "simple sentence") as a way to resolve a logical contradiction. This decision corresponds to the example dilemma we presented at the start of this article. The initial belief state, defined by a conditional and a simple sentence, can be expanded by the application of a deductive inference rule. In our paradigm, it is this resulting inferred belief that is subsequently contradicted. Because we know that human deduction is influenced by the particular form of the inference rule used (cf. Evans, Newstead, & Bryne, 1993) we are secondarily interested in whether the inference rule used in defining the initial belief set impact a subsequent belief revision rule.  While these particular experiments are not designed to discriminate between proof theoretic or mental-models theories of deduction, such evidence is relevant to expanding either of these performance models of human reasoning to embrace aspects of resolving contradiction. The final two studies examine more directly various of the alternative model-theoretic definitions of minimal change, and investigate whether minimal change—by any of these definitions—is a principle for human belief revision. This organization notwithstanding, we note that these issues—the syntactic versus model-theoretic distinction, epistemic entrenchment, and minimal change—are tightly interwoven and they bear on each experiment in some way.

## Entrenchment of Conditionals

In the first three experiments we report, we used two problem types that differed in whether the initial belief state included a conclusion drawn by the application of a modus ponens inference rule or by the application of a modus tollens inference rule. Modus ponens is the inference rule that from *If p then q,* and furthermore *p*, then infer *q.* The modus ponens belief set consisted of a conditional, the conditional's antecedent, and the derived consequent. Modus tollens is the rule that from *If p then q,* and furthermore *~q,* infer *~p*. The initial modus tollens belief set consisted of a conditional, the negation of its consequent, and the derived negation of the antecedent. We introduced contradiction with the initial belief state by providing new information—the *expansion* information—which contradicted whatever the derived conclusion was. In the modus ponens case, the expansion was *~q*. In the modus tollens case, the expansion was *p*.[6]

We defined belief-change problems using these well-studied problem types, both to provide a baseline for understanding the role of syntactic form in belief-change problems, and to make contact with existing data and theories about human performance on these classic deductive forms in a different problem context. If a conditional enjoys some kind of entrenchment by virtue of its syntactic form, people should prefer a revision that retained the conditional but reversed the truth status of the simple sentence that permitted the (subsequently contradicted) inferred sentence. A related question is whether this belief revision choice is made differently, depending on whether the belief state consisted of a modus ponens or modus tollens inference. From an AI model-theoretic viewpoint, modus ponens and modus tollens are just two different sides of the same coin: they differ only in their syntactic expression. Classical AI model-theoretic approaches would consider a revision that denied the conditional to be a more minimal change.[7] From a psychological viewpoint, it is well documented (e.g., see the survey by Evans, Newstead, and Byrne, 1993) that people find making a modus tollens inference more difficult than making a modus ponens inference. In this work, we did not want this feature of reasoning to come into play. Therefore, we provided the inferences explicitly in defining the initial belief set, and then asked whether the deductive rule used to derive them affects the belief revision choice.

The existing literature on human reasoning performance also indicates an influence of domain-specific content on the kinds of inferences that people are able or likely to draw. To account for these effects, theories have proposed the use of abstract reasoning schemas (Cheng & Holyoak, 1989; Cheng, Holyoak, Nisbett, & Oliver, 1993) and a reasoning by analogy approach (Cox & Griggs, 1982). For these initial investigations of belief-revision choices, we were not interested in investigating the direct applicability of these theoretical distinctions to the issue of belief-revision, but rather considered the general empirical findings that people reason differently with familiar topics than they sometimes do when given problems involving abstract symbols and terms. If belief revision is viewed less as a decision task driven by notions like minimal change and more as a problem of creating consistent explanations of past and current data, then we might expect the pattern of revision choices to be different when the problem content is more “real-worldly” than abstract. So, these experiments used both abstract problems (containing letters and nonsense syllables to stand for antecedents and consequents) and equivalent versions using natural language formats.

## Experiment 1

### Method

Problem Set. Table 1 gives the schematic versions of the two problem types used in this experiment. Each problem consisted of an initial sentence set, expansion information and then three alternative revision choices. The initial sentence set was labeled "the well-established knowledge at time 1." The expansion information was introduced with the phrase, "By time 2, knowledge had increased to include the following." [8] Each revision alternative was called a "theory" and consisted of statements labeled "Believe", "Disbelieve" or "Undecided About." A theory could have statements of all these types, or of just some of these types. The task for subjects was to choose one of the alternative revision theories as their preferred belief state change.

-------------------------------

Insert Table 1 about here

-------------------------------

For the modus ponens and modus tollens problems, the original sentence set included a conditional of the form $p \rightarrow q$ and either the antecedent $p$ or the negated consequent $\sim q$, respectively. In both cases, the derived inferences were included in the initial set ($q$ for modus ponens, $\sim p$ for modus tollens). The expansion information for both problems contradicted the derived inference and this was explicitly noted to subjects in the presentation of the problem. Revision choices 1 and 3 offered two different logically consistent ways to reconcile this: deny the conditional (choice 3) or retain the conditional but reverse the truth status of the simple sentence that permitted the inference (choice 1). Revision choice 2 was included to provide a choice that was non-minimal by almost any standard: it included the expansion information, denied the conditional, and labeled the simple sentence that permitted the inference to be made as “uncertain” (signified by a ? in Table 1). Note that all revision alternatives indicated that the expansion information must be believed.

Problems had one of two presentation forms: a symbolic form, using letters and nonsense syllables, and a science-fiction form. An "outer space exploration" cover story was used to introduce the science-fiction forms. Here is an example of how a modus tollens problem appeared in the science fiction condition:

On Monday, you know the following are true:

If an ancient ruin has a protective force field, then it is inhabited by the aliens called Pylons.

The tallest ancient ruin is not inhabited by Pylons.

Therefore, the tallest ancient ruin does not have a protective force field.

On Tuesday, you then learn:

The tallest ancient ruin does have a protective force field.

The Tuesday information conflicts with what was known to be true on Monday. Which of the following do you think should be believed at this point?

A corresponding symbol version of this problem was: If Lex's have a P, then they also have an R. Max is a Lex that has a P. Therefore, Max has an R. The expansion information was Max does not have an R.

Design. All subjects solved both modus ponens and modus tollens problem types. Presentation form (symbolic versus science-fiction) was a between-subjects factor. The science-fiction cover stories used several different clauses to instantiate the problems. The clauses used for each problem type are shown in Appendix A.

Subjects**.** One-hundred twenty subjects from the University of Alberta Psychology Department subject pool participated in the study. Equal numbers of subjects were randomly assigned to the symbol and science fiction conditions.

Procedure. The modus ponens and modus tollens belief problems appeared as part of a larger set of belief revision problems. The order of revision alternatives for each problem was counterbalanced across subjects. Below are excerpts from the instructions, to clarify how we presented this task to our subjects:

> ....The first part of the problem gives an initial set of knowledge that was true and well-established at time 1 (that is, some point in time). There were no mistakes at that time. The second part of the problem presents additional knowledge about the world that has come to light at time 2 (some later time). This knowledge is also

> true and well-established.... The world is still the same but what has happened is that knowledge about the world has increased....After the additional knowledge is presented, the problem gives two or more possible "theories" that reconcile the initial knowledge and the additional knowledge....Your task is to consider the time 1 and time 2 knowledge, and then select the theory that you think is the best way to reconcile all the knowledge.

## Results

Each subject contributed one revision-type choice for each of the two problem types. This gives us frequency data for how often each revision choice was selected, as a function of two variables: problem form (modus ponens v. modus tollens) and presentation form (science-fiction v. symbolic). Table 2 presents this data as the percentages of subjects choosing a particular revision choice.

From the schematic versions of the problems in Table 1, it is clear that the three belief revision alternatives for the modus ponens (MP) and modus tollens (MT) problems have a certain symmetry, even though the actual details of each revision are necessarily different. In Table 2's presentation of the data, we re-label these revision alternatives in a more general form that reflects this symmetry. For both problem types, revision choice 1 retains the conditional but reverses the truth status for the non-conditional that was the other initial belief. (For the MP problem, the expansion was $\underline{p}$, so $\underline{q}$ was the initial non-conditional. For the MT problem, the expansion mentioned $\underline{q}$; so $\underline{p}$ was the initial non-conditional.) In revision choice 2, the conditional is disbelieved and non-conditional is uncertain. Under revision choice 3, the conditional is disbelieved and non-conditional retains whatever truth value it had initially.

-------------------------------

Insert Table 2 about here

-------------------------------

In general, subjects preferred revisions in which the $\underline{p \rightarrow q}$ rule was disbelieved (revisions 2 and 3). Collapsing across presentation condition, the clearest difference between the MP and MT belief-change problems concerned which of these two rule-denial revisions subjects preferred: on MP problems, the preferred belief change saw subjects preferring simply to disbelieve only the rule; on MT problems, the preferred revision was to disbelieve the rule and to regard the initial non-conditional, $\underline{\sim q}$, as uncertain.

To analyze this frequency data, one could create a set of two-way tables for each level of each variable of interest to assess whether the distribution of frequencies is different, and compute a chi-square test of independence for each subtable; however, this does not provide estimates of the effects of variables on each other. Loglinear models are useful for uncovering the relationships between a dependent variable and multiple independent variables for frequency data. A likelihood-ratio chi-square can be used to test how well a particular model's prediction of cell frequencies matches the observed cell frequencies.

We can first ask whether the three revision alternatives were selected with equal probability, when collapsed across all conditions. The observed percentages of 22.2%, 39.9%, and 37.9% for revision choices 1, 2, and 3, respectively, were significantly different from the expected percentages ($\chi^2$=13.27, df=2, p=.001). By examining the residuals, we can identify patterns of deviation from the model. The two deviations in this case were the percentage of revision 1 and revision 2 choices.

To test whether revision choice is independent of problem type and presentation mode, we fit a model that included simple main effects for each factor, but no interaction terms. The chi-square value indicates that such an independence model does not fit the data well ($\chi^2$=15.33, p=.004, df=4). Models that included only one interaction term for revision by problem type, or only one for revision by presentation mode, were also poor fits to the observe data ($\chi^2$=12.02 and 10.52, respectively, df's=4, p's < .05). The simplest model whose predicted frequencies were not significantly different from observed frequencies included both a revision by problem-type and a revision by presentation-mode interaction term ($\chi^2$=3.18, df=2, p=.203).[9]

The means in Table 2 indicate that the pattern of difference between MP and MT choices is primarily due to differences in responses on the science-fiction problems. 58% of the science-fiction condition subjects chose to disbelieve p—>q on modus ponens belief states, while only 29% did so in the modus tollens case. The most frequently-chosen revision (54%) for a science fiction MT belief-revision was a non-minimal change: disbelieving p—> q and changing q's initial truth status from false to uncertain. Only 29% of the subjects choose this revision on the modus ponens belief state.

## Experiment 2

In Experiment 1, subjects may have been evaluating merely whether each revision option was logically consistent, independently of what the initial sentence set and expansion information was. Only two of the revisions alternatives offered minimal-changes to the initial sentence set, and this might have accounted for the close pattern of responses between symbolic-form MT and MP problems. Asking subjects to generate, rather than select, a revision would most directly address this possibility, but for these studies, we decided to retain the selection paradigm and to increase the alternatives. For Experiment 2, we included an extra non-minimal change revision and a revision in which the sentences were logically inconsistent.

### Method

Problem Set and Design. Table 3 presents the response alternatives for the modus ponens and modus tollens problems used in Experiment 2. The first three response choices were the same as those used in Experiment 1. The fourth choice denies both the rule and changes the original truth status of the initial non-conditional sentence. This is a non-minimal change and results in an inconsistent set of sentences as well. The fifth revision choice labels both the conditional and the non-conditional from the initial belief set as uncertain. These changes too are non-minimal, but the final belief set is logically consistent.

--------------------------------

Insert Table 3 about here

--------------------------------

Subjects and Procedure. Forty-three subjects participated as part of a course requirement for an introductory psychology course. All subjects solved both MP and MT problems, as part of a larger set of belief-revision problems. Only symbolic forms of the problems were used in this follow-up experiment. The instructions were the same as those used for Experiment 1.

Results

The percentage of subjects choosing each revision choice are also given in Table 3. There is some consistency in the patterns of responses across both Experiments 1 and 2. The frequency of revisions in which the initial non-conditional's truth value was changed (revision choice 1) was still relatively low (about 25%) on both problem types, as we had found in Experiment 1. About 33% of the subjects opted simply to disbelieve the conditional (revision 3) on the MP problem (as they had in Experiment 1). However, on the MT problem, changing both the conditional and the initial simple sentence to uncertain (revision 5) accounted for most of the choices.  A simple chi-square computed on the revision-choice by problem-type frequency table confirmed there was a different pattern of revision choices for these modus ponens and modus tollens problems ($\chi^2$=15.33, df=4, p=.004)

## Experiment 3

In the first experiments, we explicitly included the derived consequences in the modus ponens and modus tollens problems. In Experiment 3, we tested whether or not this inclusion of consequences as explicit elements of the initial belief set (versus allowing the subjects to draw their own conclusions) would affect revision choice. Consider, for example, problem type 1 in Table 4. This problem's initial belief set supports a simple modus ponens inference from a conditional m & d → g and the simple sentences m and d to generate the conclusion g. As in the previous experiments, there were two logically consistent ways to reconcile the ~g expansion information: deny the conditional, or deny one or more of the simple sentences that comprise the conditional's antecedent. The two revision choices reflect these two choices. Alternative 1 disbelieves the conditional and retains belief in the simple sentences; alternative 2 retains belief in the conditional and calls into question one or both of the simple sentences.

------------------------

Insert Table 4 here

-----------------------

Whether or not the initial sentence set includes derived consequences can have more profound implications when the initial belief set supports a chain of inferences. Consider problem type 2 in Table 4, in which the initial belief state is {c→h, h→m, c} and the expansion information is {~h}. One conclusion supported in the initial belief set is h. And this is in conflict with the expansion information. There are two ways to resolve this conflict: deny the conditional c→h , arriving at the final belief set of {c, h→m, ~h}. Or deny c and retain the conditional c→h, to obtain the revised belief set {c→h, h→m, ~c,~h}. Note that m cannot be inferred from either of these two revised belief states, but note also that it was a consequence of the initial belief set. Should we continue to believe in m? We can do that only if we believed in m in the first place, that is, if we drew m as a logical consequence of the first set of sentences. Otherwise, its status would be uncertain—neither believed nor disbelieved. Belief revision alternatives were provided for both these possibilities and this was investigated both in the case where logical consequences of beliefs were explicitly included in the initial belief set (as in Experiments 1 and 2) and also without explicit inclusion.

A second factor we considered in this follow-up was whether the conditional sentences in the initial belief set were propositional sentences or were universally-quantified sentences. The belief revision problem hinges on the reconciliation of conflicting information, but how that reconciliation proceeds may depend on whether it contradicts what is believed about a class (hence, is a factor relevant to predicate logic), versus what is believed about an individual (and hence is a feature of propositional logic). Therefore, we manipulated whether the initial belief set was specified by universally quantified sentences or propositional sentences for each of the problems studied in Experiment 3.

## Method

Problem Set and Design. The schematic versions of the two problem types given in Table 4 were used to create 8 different problems. Two factors were crossed for both problem types 1 and 2. The first factor was whether the minimal logical consequences of the initial sentence set were explicitly given as part of the initial belief set. In Table 4, the bracketed simple sentences were explicitly listed as part of the initial belief set in the consequences-given condition or were omitted in the no-consequences given condition.

The second factor, sentence-form, was whether the belief set was based only on propositional sentences, or concerned sentences about universally-quantified arguments. Thus, one propositional form of a conditional was If Carol is in Chicago, then she stays at the Hilton Hotel, while the universally-quantified form was Whenever any manager from your company is in Chicago, s/he stays at the Hilton Hotel. The associated simple sentences in each case referenced a particular individual. For the propositional example, the sentence instantiating the antecedent was *You know that Carol is in Chicago.* For the universally-quantified condition, it was *You know that Carol, one of the company managers, is in Chicago.*

For problem type 1, the revision choices were either to disbelieve the conditional (revision alternative 1) or to disbelieve one or both of the initial simple sentences (revision alternative 2). The same distinction holds for problem type 2, which had four revision alternatives: alternatives 1 and 3 involved denying the conditional *c→h*, while revision choices 2 and 4 retained the conditional and instead changed *c* to *~c*. The other key distinction in problem type 2's revision alternatives concerned the status of *m*, which was the chained inference that the initial belief set supports. Revision choices 1 and 2 labeled *m* as uncertain; revision alternatives choices 3 and 4 retained *m* as a belief.

All of the problems were presented in natural language formats. The following text illustrates how Problem Type 1 appeared in the consequences given—propositional condition:

> Suppose you are reviewing the procedures for the Photography Club at a nearby university, and you know that the following principle holds:
>
> > If the Photography Club receives funding from student fees and it also charges membership dues, then it admits non-student members.
>
> You further know that the Photography Club does receive funding from student fees. It also charges membership dues. So you conclude it admits non-student members.
>
> You ask the Photography Club for a copy of its by-laws and you discover
>
> > The Photography Club does *not* admit non-student members—all members must be registered students.

Subjects and Procedure. Thirty-five University of Alberta students served as subjects, to fulfill a course requirement for experiment participation. Problems were presented in booklet form, which included other belief-revision problems as fillers. All subjects solved all four versions of both problem types 1 and 2: no consequence—propositional, consequences given—propositional, no consequences—quantified, consequences given—quantified. There were six pseudo-random orders for the problems within the booklet; within each order, the four versions of any given problem were separated by at least two other problems of a different type. The order of response alternatives for each problem was also randomized.

## Results

For problem type 1, revision choice 1 (disbelieving the conditional; see Table 4) accounted for 82% of the revision choices. This is consistent with the pattern of choices in Experiment 1 results on science-fiction problems and this preference to disbelieve the conditional was not affected by whether or not the modus ponens inference was explicitly listed in the initial sentence set nor by the use of propositional v. universally-quantified sentences. In terms of the first factor, we note that people generally find modus ponens an easy inference to make, and these results confirm that the general preference to disbelieve the conditional does not rest on whether the contradicted inference is explicitly provided. Concerning propositional v. universally-quantified sentences, we observe that it is difficult to construct if p —>q sentences that are not, somehow, interpretable as universally-quantified over time. Thus, even sentences like If Carol is in Chicago, then Carol is at the Hilton, may be interpreted as For all times when Carol is in Chicago, ..... There seems to be little in the line of systematic, empirical study of the effect of propositional v. single quantifier v. multiple quantifier logic upon people's reasoning (although both Rips, 1994, Chapts. 6 and 7, and Johnson-Laird & Byrne, 1991, Chapts. 6 and 7, address this issue in their respective computational frameworks). Nonetheless, it seems clearly to be an important issue for studies that place an emphasis upon recognition of contradictions, since the impact of contradictory information upon "rules" is different in these different realms.

There was also no impact of either the consequences-given or the sentence-form factor on the patterns of revision choices for Problem Type 2, in which the initial belief set contained an intermediate conclusion $h$ and then a chained conclusion $m$, that depended on $h$, and where expansion information contradicted $h$. The percentage of revision choice 1 (denying the conditional $c \longrightarrow h$) accounted for 52% of the choices; choice 2 (denying the non-conditional sentence $c$) accounted for 29% of the choices. In both these cases, the status $m$, the chained inference that depended on $h$, was labeled uncertain. Revision alternatives 3 and 4, which were analogous to alternatives 1 and 2 except that they retained belief in $m$, accounted for 14% and 5%, respectively, of the remaining choices. The preference to change $m$'s truth status from true to uncertain rather than retain it as true is interesting: it is an additional change to the initial belief state beyond what is necessary to resolve the contradiction. Perhaps people's revision strategy is guided more by the recognition that a belief depends on another than upon minimizing the number of truth values that change from one state to the next.

## Discussion

In Experiments 1-3, we aimed to identify what kinds of revision choices subjects would make in symbolic and non-symbolic types of problems, with the former providing some kind of baseline for whether a conditional statement enjoys some level of entrenchment merely as a function of its syntactic form. Our second concern was to assess whether belief revision choices were affected by the composition of an initial belief set, i.e., whether it was defined through the use of the conditional in a modus ponens or modus tollens inference. This offers us a bridge between belief revision (as a task of making a deliberate change in what is to be "believed" in the face of contradictory information) and the data and theories on deductive reasoning.

There was no evidence that people preferred to entrench the conditional on these tasks. In the choices we gave subjects, there was one way to continue to believe the conditional and two ways to disbelieve it. If people were equally likely to retain the conditional as they were to abandon it, we might expect 50% of the choices falling into the keep-the-conditional revision, with the two ways to disbelieve it each garnering 25% of the choices. On the symbolic problems in Experiments 1 and 2, the frequency of retaining the conditional after the expansion information was only about 25% on both modus ponens and modus tollens problems; it was even lower on the natural language problems.

Although subjects' preference was to abandon belief in the conditional, the way in which this occurred on modus ponens and modus tollens problems was slightly different. On modus ponens problems, subjects disbelieved the conditional but continued to believe the non-conditional sentence as it was specified in the initial belief set. On modus tollens problems, subjects tended towards more "uncertainty" in the new belief state: either denying the conditional and deciding the non-conditional was uncertain (Experiment 1), or labeling both as uncertain when that was an option (Experiment 2). These tendencies on modus tollens problems could be interpreted as conservative revision decisions, since neither the initial conditional nor the initial non-conditional sentence is explicitly denied; on the other hand, they correspond to maximal changes because the truth values of both initial beliefs are altered. We leave further discussion of entrenchment issues to the General Discussion.

It is natural at this point to consider the relationship between this belief-change task and standard deduction, and to ask whether this task and its results can be understood as a deduction task in some other guise. There are two reasons we think it is not. First we consider the task demands and results for the modus ponens and modus tollens belief revision problems, and then briefly outline results we have obtained on belief expansion problems that did not involve a contradiction.

### The task demands of the modus ponens and modus tollens belief-revision problems

We can neutrally rephrase the modus ponens belief-change problem that subjects faced as "Make sense of $[p \rightarrow q, p, q] + [\sim q]$," where the first sentence set represents the initial belief set and the second signifies the expansion information. Since subjects had to accept the expansion information, what we call the modus ponens problem thus becomes "Make sense of $[p \rightarrow q, p, \sim q]$, such that $\sim q$ is retained." Similarly, the modus tollens problem is "Make sense of $[p \rightarrow q, \sim q, p]$, such that $p$ is retained." Because these two problems are semantically equivalent, the forms in the set of propositions to be considered are the same and the models of these sentence sets are the same. The difference lies only in the nature of the derivation in the initial sentence set, and the corresponding constraint on what must be retained after the revision.

What we have called the modus ponens belief revision problem could be construed as an modus tollens deduction problem, if subjects consider only the conditional in combination with the expansion information: "Given $[p \rightarrow q] + [\sim q]$, what can I derive?" The invited modus tollens inference is $\sim p$. If they derived this, they could at least consider retaining the conditional and changing $p$ to $\sim p$ in their belief-state change. The trouble that modus tollens inferences present for people could in this way explain the observed prevalence of disbelieving the conditional on modus ponens belief revision problems.

Applying this same perspective on the task to the modus tollens problem, we would see the modus tollens belief revision problem becoming an modus ponens deduction problem, if only the conditional and the expansion information are considered: "Given $[p \rightarrow q] + [p]$, what can I derive?" People have little difficulty with modus ponens and under this analysis, it would be an "easy inference" to conclude $q$, and so be led to reverse the truth status of $\sim q$ as the belief change. But the majority of subjects did not do this—on these problems as well, they disbelieved the conditional. Therefore, it does not seem that our general pattern of disbelieving the conditional in belief revision can be reduced to, and accounted for by, the nature of the difficulties in making certain types of standard deductive inferences.

It is possible that subjects did not accept the modus tollens belief set as consistent in the first place. (People have difficulty both in generating modus tollens inferences and in validating them when they are provided—cf. Evans, Newstead, & Byrne (1993), p.36). So perhaps this could be used to account for why there was high percentage of "everything but the expansion information is uncertain" revisions on modus tollens problems in Experiment 2. However, this does not account for why, on these modus tollens problems, subjects would not simply focus on both the conditional and the expansion information, and then draw an modus ponens inference—that would lead to changing the truth status of the initial simple sentence, as opposed to what they in fact did.

### Deductive reasoning and belief-state expansions

The second reason we believe these tasks are not reducible to equivalent deductive reasoning problem stems from results we obtained on other belief-state expansion problems, in which the expansion information did not contradict the initial belief set (Elio & Pelletier, 1994). These problems used two different but logically equivalent forms of a biconditional: $p$ *if and only if* $q$ and $( (p \,\&\, q) \vee (\sim p \,\&\, \sim q) )$. The expansion information was sometimes $p$ and at other times $\sim p$. Unlike the belief revision problems, these problems have a deductively "correct" answer: given $p \leftrightarrow q$ (in either form) as an initial belief, with the sentence $p$ as the expansion, it logically follows that $q$ should be asserted and made part of the belief state. (And if $\sim q$ is the expansion, then $\sim p$ should be believed). If we view the biconditional-plus-expansion information problems as biconditional modus ponens (or biconditional modus tollens) problems, then we would expect that subjects presented with our biconditional and disjunctive belief expansion problems should behave like the subjects given biconditional and disjunctive deductive problems in other studies. Yet we found that subjects asserted *q* on the $p \leftrightarrow q$ form of our biconditionals much less frequently (about 72%) than typically reported for these problems presented as standard deduction tasks (e.g., 98% accuracy in Johnson-Laird et al., 1992). And fully 56% of subjects given the biconditional in disjunctive form followed by the belief expansion *p* did not augment their belief set with *q*, when the problem was presented with a science-fiction cover story. Instead, they decided *q* was uncertain and that the biconditional itself was uncertain or unbelievable.

In sum, we believe that the task of belief revision, even in the relatively constrained way we have defined it here, does not simply unpack into deductive reasoning, particularly when natural-language formats are used for the problem. That is, subjects may not integrate information arriving across time (e.g., learning "later" that *p* holds true) into a belief set in the same way as information known to be true at the same time ("From *If p is now true, then q is also true*, and furthermore *p is now true*, what follows?"). It may be that the belief revision task invites the reasoner to make certain assumptions about evidence that is not explicitly included in the initial or subsequent information; it may also be that couching the task as changes in beliefs invites a more conservative strategy than what characterizes people's choices on formal logic problems.

## On models of belief states and deduction

The experiments we designed do not speak to whether belief states are best modeled as sets of sentences or sets of models. However, we can observe the following. First, AI competence models are typically not concerned with human performance, yet they sometimes appeal to human rationality to justify their particular perspective. For example, a syntax-based competence model proponent may point to the fact that a model-based perspective involves an infinite number of models, when taken to the extreme; and because that is so clearly beyond the capability of human cognition, such modeling cannot be appropriate. A model-theoretic proponent might say that it is only via models of the actual world that the meaning of the sentences has any reality. Even acknowledging that competence models are *not* likely to be interested in belief revision decisions on modus ponens and modus tollens based belief states; we *can* nonetheless say that a model-theoretic competence framework could never model any of these kinds of differences, since modus ponens and modus tollens are indistinguishable from the perspective of formal model theories. Further, our finding that people seem to prefer to abandon the conditional is problematic for model-theoretic frameworks, unless they retain some mapping between each sentence and the model which that sentence generates. But there are difficulties for a syntactic-based perspective. It is unclear that the syntactic form of sentences *per se* should be a primary tag for guiding belief revision decisions. Indeed, our finding that people were more willing to abandon the conditional on natural language problems than on symbolic problems suggests that there are other, non-syntactic considerations at play that may serve as pragmatic belief revision principles. We return to this issue in the General Discussion.

The belief revision results we obtained do not speak directly to performance theories of human deduction, but there are some important observations we can make here as well. First, the Johnson-Laird mental models framework could possibly accommodate the general preference to deny the conditional, by the preference ordering it puts on models that different types of sentences generate. The mental model of $p \rightarrow q$ is "[$p$ $q$]..." where [$p$ $q$] represents the initial explicit model, in which both $p$ and $q$ are true, and the ellipsis “...” represents that there are additional models of this sentence (corresponding to possible models in which $p$ is not true; Johnson-Laird, Byrne, and Schaeken, 1992). For our modus ponens problem, the initial sentence set is $p \rightarrow q$, $p$, and $\therefore q$. Let C indicate models of the conditional, and S to indicate models of simple sentences in the initial belief set. Hence, the initial modus ponens model set would be C: [$p$ $q$]..., S: [$p$], S:[$q$], respectively. Note that the models for the simple sentences are consistent with what the mental models theory proposes as the initial explicit model for the conditional. The modus ponens expansion information is ~q and we denote its model as E:[~q]. Suppose a subject compares the expansion model E:[~q], which must be retained in any revision, to each of the models from the initial set. The expansion model would eliminate the model S:[q], be silent on the model S:[p], and eliminate the model C:[p q] of the conditional. By this process, the preferred revision choice should be to deny this model of the conditional and the retain the non-conditional sentence p. In fact, this choice accounted for 75% of the modus ponens revisions in Experiment 1 and about 60% in Experiment 2. By the same general reasoning, the mental-models approach would find itself predicting a preponderance of conditional denials for modus tollens problems. While we did find this is true in general, there would have to be some further account for people's greater tendency to decide the conditionals are uncertain (rather than false) on modus tollens problems than on modus ponens problems.

From a proof-theoretic perspective, Rips (1994, pp. 58-62) directly considers the problem of belief revision as the issue of which of several premises to abandon in the face of contradiction, acknowledging that deduction rules cannot alone "solve" the belief revision problem. He discusses a multi-layer approach, in which the principles governing belief revision decisions are themselves "logic-based processing rules" that co-exist with the deduction rules that he proposes as components of reasoning and problem-solving. Thus, a proof-theoretic approach might be extended to deal with our belief revision results by having an explicit higher-level rule that, when contradiction is recognized, indicates the action of disbelieving a conditional form when it is one of the premises. But even without an appeal to this approach, it is possible to consider a proof-theoretic account of our results, as we did for the mental-models perspective, using Rips' (1994) framework. Recall again the above perspective that portrayed the modus ponens belief-revision problem as boiling down to " Given [p—>q, p] + [~q] and the constraint that ~q must be retained as a belief, what can you prove?" One can imagine that a subject formulates two competing sets of premises. One set is [p—>q, ~q]. There is no direct modus tollens rule in Rips' theory (the modus tollens inference is accomplished through the application of two other inference rules), thus accounting for the notion that modus tollens proof for ~p is difficult and may halt. On the other hand, there is a readily available inference rule ("and introduction") that can apply to the other combination of premises [p, ~q] to yield [~p and q]. From this perspective, subjects might reach a state that they can more easily recognize as valid and that may be why they prefer a revision in which these sentences are retained and the conditional is disbelieved. On the modus tollens problem, we can characterize the belief revision dilemma as "Given [p—>q, ~q] + [p] and the constraint that p must be retained, what can you prove?" The modus ponens rule is readily available according to Rips' theory, and so the premise combination [p—>q, p] easily yields q. Just as easily, the other combination of premises [~q, p] yields [p and ~q]. The greater tendency to prefer revisions that label the conditional (and the non-conditional) "uncertain" in the modus tollens belief revision case relative to the modus ponens belief-revision case may reflect subjects' ability to prove something from both combinations of premises (as we have stated them) and their appreciation that they have no reason to prefer the premises of one proof over the other in these simple problems.

Our goal in considering how two contrasting perspectives of deductive reasoning might accommodate our results was not to support one over the other; neither was it our motivating intent. The accounts we sketched above are offered as speculations on how each perspective might be extended into the realm of belief revision, given their representation and processing assumptions about deductive reasoning. Such extensions are an important component for an integrated theory of reasoning and required much more consideration than we have briefly allowed here.

## Models and Minimal Change

As we noted earlier, one of the desiderata of the classical AI belief revision perspective is that an agent should make a minimal change to its initial belief set, when resolving any conflict that results from new information. Within a syntactic approach, the definition of change is computed from the number of formulas retained from one belief state to another; there are not many different ways to compute this number, since the formulas are fixed. The primary issue is whether or not the set of formula is closed, i.e., includes all consequences of the initially-specified set of sentences. When the set of formulae is not closed, making a formula become part of the explicit belief set is regarded as more of a change than having it be in the implicit beliefs.

Within a model-theoretic approach, it turns out there is more than one way to compute what a minimal change might be, even for the simplest problems. In this section, we present the gist of some alternative computational definitions of minimal change. None of these approaches were devised as psychological models of how humans might manipulate alternative models in the face of conflicting information. And while the ways the algorithms that compute minimal change might not be psychologically plausible, the final change that each one deems minimal often corresponds to an intuitively reasonable way of integrating both the old and new belief information. We provide simple algorithmic interpretations of each of these minimal change definitions in Table 5 and highlight the functional effects of computing minimal change according to one algorithm or another.

A straightforward way to quantify the degree of change is to count the number of propositions whose truth values change if one model (e.g., expansion information) is integrated with another model (e.g., the initial belief set). The tricky part comes when there is more than one model of the initial belief set, or of the expansion information, or both. Clearly, there will be more than one possible interpretation for a sentence set whenever there is an explicit uncertainty. By explicit uncertainty, we mean a belief sentence that directly mentions that the truth status of some proposition is either true or false. Hence, in the sentence set (p, q ∨ ~q), q is explicitly uncertain, so there are two models of this sentence set: [p ~q], [p q]. Suppose, however, that the initial sentence set were "Either p and q are true at the same time, or they are false at the same time" and that the expansion information is "p is false, q is true, and furthermore r is true." The initial belief state has two models, [p q], [~p ~q], and both p and q are explicitly uncertain. The proposition r was not in either of the initial models of the world. But clearly, its truth status (along with every other possible sentence) in the initial belief set was, in hindsight, uncertain. This is what we call implicit uncertainty, and all the algorithms in Table 5 construct different models of the initial belief set to accommodate the implicit uncertainty about r just as if it were explicitly uncertain in the first place. Thus, the computations for minimal change for this problem would begin with these models of the initial belief set [pq~r], [pqr], [~p~q~r], and [~p~q~r]. As we shall see in the first example below, this same approach of creating extra models also applies when a sentence that is present in the initial belief set is not mentioned the expansion information.

One approach to determining a minimal change is to chose a model of the expansion sentences that is the minimal distance from some model of the initial belief set. Suppose an initial belief is "Either p, q, r, s are all true at the same time, or they are all false at the same time." So there are two different models of this initial belief: [p q r s] and [~p ~q ~r ~s]. Expansion information such as "p is true, s is false, and r is false" contradicts this initial belief state and furthermore does not mention anything about q. There are then two models of the expansion, one in which q is true [p q ~r ~s] and one in which it is false [p ~q ~r ~s]. The latter model of the expansion is "close" to the second model (disjunct) of the initial belief set and is indeed "closer" than either expansion model is to the first model of the initial belief set. By this reasoning, a new belief state that represents a minimal change on the initial state is [p ~q ~r ~s]). This is the gist of the minimal change approach proposed by Dalal (1988) and summarized as Algorithm D in Table 5. More formally, Dalal's revision of a belief set by an expansion sentence is a set of minimal models where (a) each member of this set satisfies the expansion information, and (b) there is no other model of the initial belief set that also satisfies the expansion information and differs from any model of initial belief set by fewer atoms than the set of minimal models. The revision results in the set of all these minimal models. Thus, Dalal's algorithm settles on one model of the expansion information, if possible, and in doing so, can be viewed as retroactively settling on one particular model of the initial belief set.

An alternative intuition would hold that: only informative (non-tautological) initial beliefs can be used to choose among multiple interpretations of the expansion information, if they exist. This is one way to interpret an algorithm proposed by Weber (1986). Simply put, Weber's algorithm first identifies the initially-believed sentences that must take on whatever truth values are specified for them in the expansion. For the same example in the preceding paragraph, this set would contain the sentences p, r, and s, because they each have a specific value they are required to take, according to the new information. These sentences are then eliminated from the initial belief set to identify what (if any) informative sentences propositions might be retained from the initial belief set. Subtracting p, r, and s from the initial belief set {[p q r s], [~p ~q ~r ~s]}leaves [q v ~q], which is a tautology, and by Weber's algorithm, leaves no (informative) proposition. (Had there been some other sentence which was in both of the initial models, it would have then been assigned to the revised belief state). The algorithm then conjoins these two components: the truth values of p , r, and s as determined by the expansion information [p ~r ~s] and whatever can be retained with certainty from the initial belief set, which here is the empty model [ ]. Whereas Dalal's revision for this problem would be [p ~r ~s ~q], Weber's minimal revision would be [p ~r ~s], with q implicitly uncertain by virtue of its absence from the model. A simple algorithm that corresponds to this approach is given as Algorithm W in Table 5.

------------------------

Insert Table 5 about here

------------------------

Borgida (1985) proposes an algorithm that is similar to Dalal's, but produces what might be considered a more conservative belief-state change. Essentially, each expansion model is compared to each initial belief-set model: the expansion model that produces a minimal change for a particular initial-belief interpretation is remembered. All these expansions that are minimal with respect to some model of the initial belief set are then used to define the new belief set. An algorithm that captures this approach is given as Algorithm B in Table 5. Consider a case where there is more than one interpretation of the initial belief set. If [p q ~s] is the initial belief set, and [~p ~q r ~s] and [~p ~q ~r s] are two models of the expansion information, then two models of the initial belief set are considered: the first contains r and second contains ~r. Both interpretations of the expansion information define a minimal change with one of the interpretations of the initial belief set (the first expansion disjunct with the first interpretation of the belief set, and the second expansion disjunct with the second interpretation of the belief set). Thus, both [~p~q r ~s] and [~p ~q ~r s] are on the stack after step B1.2. Since neither of these is minimal with respect to the other, the final belief set consists of guaranteed truth values for those propositions on which the interpretations agree and uncertain truth values for propositions on which they disagree, yielding a final belief state of [ ~p~q {r~s ∨ ~rs}]. Algorithm B differs from Algorithm D in that each model of the initial belief set identifies, in Algorithm B, what model of the expansion information would result in a minimal change (by number of propositions changed). Once one of the expansion models is identified as minimal with respect to a particular model of the initial belief set, there is no further check of whether one change is more or less minimal than some other combination of initial-belief interpretation and expansion-interpretation (as Algorithm D does on step D2). This can be viewed as a more conservative belief-state change, because there isn't the possibility of settling on one particular model of the initial belief state.

Satoh (1988) proposed belief revision operator that is a less-restricted version of Borgida’s revision operator, when applied to the propositional case. The feature that makes it less restricted is illustrated in Algorithm S, which is identical to Algorithm B, except that step B1.2 in Algorithm B occurs outside the first control loop as step S2 in Algorithm S. Functionally, this difference means that there is no pruning of non-minimal changes with respect to a particular belief-set model (as on Step 1.2 in Algorithm B). Instead, the entire set is saved until step S2, which removes any change that subsumes another change. After S2, all changes that remain are minimal. Step S3 then finds a model of the expansion that is consistent with the minimal set of necessary changes. Put more intuitively, this algorithm crosses all interpretations of the initial belief set and all interpretations of the expansion set to create the model set from which a minimal change is computed. The functional effect is that, when there is just one model of the initial belief set, that model may "choose" the closest interpretation of the expansion information; when there is just a single version of the expansion information, that model may "choose" among alternative models of the initial information. Only the latter may occur under the Borgida algorithm.

We are not interested so much in the means by which these alternative model-based revision frameworks define minimal change, as we are in the way they capture alternative intuitions about manipulating multiple models. In Algorithm D, the way that minimal change is computed can have the effect of "selecting" one of multiple interpretations of the initial belief set. The effect of Algorithm B is to retain multiple models in the new belief set when there are multiple models of the expansion information. Algorithm S will compute a new belief state with multiple models, when multiple models exist in both the initial and expansion information; but it can use a single model of either to produce a single model of the new belief set. Finally, Algorithm W uses the expansion information to define what can be believed with certainty; other belief-set sentences not mentioned in the expansion information may decide between multiple interpretations of the expansion information, but only if their truth value was known with certainty in the first place (i.e., was true in every model or false in every model of the initial belief state).

There are plausible elements in each of these approaches for principles that might dictate how people deal with multiple interpretations of information when resolving inconsistencies. Our interest was whether which, if any of them, corresponded to how people integrate multiple models in a belief revision task. As the reader might surmise, for any particular problem, some or all of the methods could yield the same final belief set. It is possible, however, to define a set of problems for which a pattern of responses would distinguish among these alternative approaches. We developed such a problem set to obtain data on whether people follow a minimal change principle, as defined by any of these approaches. The revision problems were very simple: there were either one or two models of the initial belief set and either one or two models of the expansion information. The problem sets were designed to distinguish among the four model-based minimal change frameworks described above.

## Experiment 4

### Method

Problem Set. Table 6 gives the problem set used for Experiments 4 and 5. The first five problems in this table were used only in Experiment 4; problem 6 was added for Experiment 5. For economy of space, we write sentence letters adjacent to one another to mean ‘and’. Thus, the problem 1 notation (pqrs) v (~p~q~r~s) means “Either p, q, r, and s are each true at the same time or else they are each false at the same time.”

-----------------------

Insert Table 6 about here

-----------------------

The subscripts for the revision choices in Table 6 correspond to the particular model-theoretic definition of minimal change: D for Algorithm D, W for Algorithm W, and so forth. Experiment 4 offered subjects two revision choices for Problems 1-5 (of Table 6); these each corresponded to one or more of the four definitions of minimal change we outlined in the previous section. It can be seen that each of the four algorithms selects a different set of answers across these five problems: Algorithm D selects answers <1,1,1,1,1> for its five answers; Algorithm B selects answers <2,2,2,1,1>; Algorithm S selects answers <2,2,1,1,2>; and Algorithm W selects answers <2,2,2,2,2>.

Design. Problem type was a within-subjects factor; all subjects solved all five problems. As in Experiment 1, presentation form (symbolic v. science-fiction stories) was manipulated as a between-subjects factor. Appendix B shows how the initial-belief sentences and the expansion sentences were phrased in the symbolic condition; the revision alternatives were phrased in a similar manner. Different letters were used in each of the problems that the subjects actually solved. The five different science-fiction cover stories were paired with the problems in six different ways.

Subjects and Procedure. The same 120 subjects who participated in Experiment 1 provided the data presented here as Experiment 4. Sixty subjects were assigned to the symbolic condition and sixty were assigned to the science-fiction condition. Equal numbers of subjects received the six different assignments of science-fiction cover stories to problems. No science-fiction cover story appeared more than once in any subject's problem booklet. Other details about the procedure and instructions were as described for Experiment 1.

## Results

Unlike the modus ponens and modus tollens belief revision problems, there was no significant effect for the symbolic versus science-fiction manipulation on these problems. Table 6 presents the percentage of subjects choosing each possible revision choice, collapsed across presentation condition. The only planned comparisons concerning these data were within-problem differences, i.e., whether one revision choice was preferred significantly more often than another. Within each problem, there is a clear preference for one revision over the other: subjects chose revisions that most closely matched the form of the expansion information. We also tabulated the number of subjects whose response pattern across problems matched the particular pattern associated with each revision algorithm described in Table 5. Virtually no subjects matched a particular response pattern for all five problems.

## Experiment 5

A concern about these data is that subjects were not following any particular model of change at all, but simply using the expansion sentence to define the new belief set. This could mean that they viewed the problem as an update, rather than a revision, problem (i.e., the world has moved to a new state defined by the expansion and there is no reason to maintain anything from the initial belief state), or it could mean that they were simply not engaged in the task.  Since the same subjects generated distinct problem-specific patterns of responses in Experiment 1, we do not believe the latter possibility holds.

In Experiment 5, we included two additional response alternatives for each problem in order to test whether subjects continued just to adopt the expansion information (which might be the simplest interpretation of the results). Revision choice 3 was a non-minimal change model that was consistent with some interpretation of the expansion information. Revision choice 4 included only those sentences whose truth values were not contradicted within the expansion information or between some model of the initial sentences and the expansion. Basically, revision choice 4 offered the minimal number of sentences that could be known with certainty and made all other conflicts between truth values become "uncertain."

We also added Problem 6, which was isomorphic in form to Problem 5, except that the initial belief set consisted of all negated sentences rather than of all positive sentences. If subjects have a bias for models that consist primarily of non-negated sentences, then they should prefer such "positive" models regardless of whether they are minimal change models. Problems 5 and 6 differed only in whether the sentences in the initial set were all true or all false. Note the symmetry between revision choices 1 and 3 for these problems: the revision [~pqr], with one negated sentence, is a minimal change model for Problem 5 but a non-minimal change model for Problem 6. Conversely, [p~q~r] is the minimal change model for Problem 6 and a non-minimal change model for Problem 5. If subjects are biased towards revisions that maximize non-negated sentences, then there should be an interaction between the form of the initial belief set and the revision selected. Finally, we stressed in the instructions that both the initial and subsequent information should be considered before determining what should or should not be believed, just in case subjects believed that the expansion information should replace the initial belief set.

## Method

Forty-three subjects solved problems 1-6 from Table 6 in random order. Since Experiment 4 had shown no effect for symbolic v. science-fiction presentation, the problems were presented in symbolic form only and the response alternatives appeared in different random orders for each subject.

## Results and Discussion

The percentages of subjects choosing each revision choice in Experiment 5 are given in Table 6. As in Experiment 4, Experiment 5's subjects did not consistently obey any particular pattern of minimal change. First, it is striking that revision choice 1 was never the most preferred revision—it is the syntactically simplest way of specifying a model that accommodates the expansion sentence and corresponds to Algorithm D, which has an intuitively simple notion of minimal change. The second feature of the results concerns the relative percentages of revision 2 (in which the new belief state is simply the adoption of the new information) and revision 4. While revision choice 2 was the clear preference in Experiment 4, it was no longer the clear favorite here. Generally speaking, if subjects were given the option of tagging certain sentences as "uncertain" (revision 4), they gravitated to this choice over a revision that more precisely (and more accurately) specifies the uncertainty as multiple models (revision 2). One conjecture is that subjects elect to use revision 4 as short-hand way of expressing the uncertainty entailed in having multiple models of the world. That is, they may see "p and q are both uncertain" as equivalent to (p~q) ∨ (~pq), although, of course, it is not. It is unclear whether subjects appreciate the 'loss of information' inherent in such a specification.

Problems 5 and 6 were of particular interest, because they differed only in whether the initial belief set consisted of positive or negated sentences; the expansion information was the same. The set of revision alternatives was also identical. As with the other problems, the most preferred revision choice (about 40%) was to declare all sentences uncertain, when their truth value differed in two different models of the expansion information (and subjects did not merely adopt the multiple model described by the expansion information as the new belief state, as they had in Experiment 4). However, if we restrict our attention just to the percentage of revision 1 and revision 3 choices in these problems, we see that about the same number of subjects (20%) chose the revision ~pqr when it served as minimal change revision 1 for problem 5 and also when it was the non-minimal revision 2 for Problem 6. Conversely, only 7% of the subjects chose p~q~r when it was the non-minimal revision 1 for Problem 5, but also only 7% chose it when it was (the minimal change) revision 3 for Problem 5. A simple chi-square computed on the response-choice by problem type (Problem 5 v. Problem 6) frequency table was marginally significant ($\chi^2$ =7.52, df=3, p =.057). These results suggest that there may be a bias against revisions that have more negated beliefs than non-negated beliefs in them. There is some suggestion of this in problem 2 as well, in which 35% of the subjects choose a non-minimal change revision (revision 3) than either of the two minimal change revisions (revisions 1 and 2). Such a finding itself is certainly consistent with body of evidence indicating that reasoning about negated sentences pose more difficulties for subjects (see, e.g., Evans, Nestead, Byrne, 1993, on "negated conclusions")**;** hence, people may prefer to entertain models of situations that contain fewer negations, when possible. This possibility of a bias against models with negations needs further, systematic study. In sum, Experiments 4 and 5 suggest that subjects are not following any single model-based minimal change metric and do not integrate the expansion information whole-heartedly. Despite the availability of choices that could be selected via simple matching procedure between disjuncts appearing in the initial and new information (revision 1 across all problems), our subjects seem to prefer belief states that consist of single models and models with non-negated beliefs, when possible.

## General Discussion

We can summarize the main findings from this study as follows. First, to resolve the inconsistency that new information creates with an existing belief set that consists of simple sentences ($\underline{p}$, $\underline{q}$) and conditional sentences ($\underline{p \rightarrow q}$), the preferred revision was to disbelieve the conditional rather than alter the truth status of one of the initial simple sentences. This preference was even stronger on problems using science-fiction or familiar topic cover stories than it was using symbolic formulas. Second, there were some differences in revision choices depending on whether the initial belief set was constructed by using a modus tollens or modus ponens inference. Subjects more often changed the truth status of the initial simple sentence (and the conditional, when there was that option) to "uncertain" on the modus tollens problems than they did on the modus ponens problems. Third, we observed that the patterns of revision choices on the simple problems we investigated does not depend on whether or not the (modus ponens) inference was explicitly listed in the initial belief set or whether subjects were left to perform the inference themselves. Fourth, we note that the patterns of revision did not change when the initial belief state was constructed from purely propositional reasoning or used universally-quantified inferences. Fifth, we discovered that when an implied conclusion of the initial belief set itself gives rise to yet another conclusion, and when the first of these conclusions is contradicted by the expansion information, then the status of the second conclusion is regarded as "uncertain." Finally, we investigated alternative model-theoretic definitions of minimal change. We found that subjects did not adhere to any of these particular prescriptions, some of which (e.g., Algorithm D) can be construed as a fairly straightforward matching strategy between a model in the initial information and a model of the expansion information. Even when the initial belief state had only one model, subjects did not use it to chose among alternative models of (uncertain) expansion information; and even when there was only a single model of expansion information, subjects did not use this to chose among alternative models of an (uncertain) initial belief state. A disjunct of multiple models can specify how the truth value of one sentence co-varies with another's; subjects did not prefer such multiple-model specifications of a belief state as a way to represent uncertainty. They instead single-model revisions that retained only sentences that had an unambiguous truth values across the initial and expansion information, and labeled all other sentences as uncertain (even though this results in a loss of information). There is a possibility as well that people prefer revisions that contain positive rather than negated sentences; this requires further study. In the remainder of this section, we consider these results for notions of epistemic entrenchment and minimal change.

## On Epistemic Entrenchment

The rationale behind a notion like epistemic entrenchment is that, practically, an agent may need to choose among alternative ways to change its beliefs, and intuitively, there will be better reasons to chose one kind of change over another. These better reasons are realized as a preference to retain or discard some types of knowledge over another; the issue is what those epistemically-based principles of entrenchment are or ought to be. As we noted in the introduction, some theorists have argued that conditional statements like $p \rightarrow q$ may warrant, *a priori,* a higher degree of entrenchment than some other sentence types, not because there is something to be preferred about material implications, but because that form often signals "law-like" or predictive relations that have explanatory power. And law-like relations, because of their explanatory power, should be retained over other types of knowledge when computing a new belief state.

We did not find evidence for this kind of entrenchment as a descriptive principle of human belief revision in the tasks we studied. In general, the frequency of continuing to believe the conditional was lower than what might be expected by chance, and lower still on natural language problems. Finding that belief-revision choices changed when the problems involved non-abstract topics is not surprising, for there are many results in the deductive problem solving literature indicating that real-world scenarios influence deductive inferences, serving either to elicit, according to some theories, general pragmatic reasoning schemas (e.g., Cheng & Holyoak, 1989) or, according to other interpretations, specific analogous cases (Cox & Griggs, 1982). On the other hand, there was no domain-specific knowledge subjects could bring to bear about a science-fiction world. Indeed, the clauses used to make science-fiction sentences are not unlike those used by Cheng and Nisbett (1993) as "arbitrary" stimuli to investigate causal interpretations of conditionals. Nonetheless it is clear that subjects revised and expanded non-symbolic belief sets differently than they did symbolic belief sets.

Subjects may have interpreted the science-fiction conditional relations as predictive, or possibly causal, relations. The instructions that set up the science-fiction problems enjoined subject to imagine that information about an alien world was being relayed from an scientific investigative team. This may have prompted a theory-formation perspective, based on the assumption that even alien worlds are governed by regularities. The generation of, and belief in, these regularities depends on observations. The initial belief set had such a regularity in it (the conditional), plus a “direct observation” sentence. When the expansion information indicated that the inference from these two was contradicted, the "denial" of the conditional is one way of asserting that the regularity it expresses, as specified, does not hold, *in this particular case*. Cheng and Nisbett (1993) found that a causal interpretation of *if p, then q* invokes assumptions of contingency, namely that the probability of $q$'s occurrence is greater in the presence of $p$ than in the absence of $p$. Subjects may have viewed the (contradictory) expansion information in the modus ponens and modus tollens problems as calling this contingency into question.  Such a perspective only makes sense when the problems are not manipulations of arbitrary symbols, and is consistent with our finding a higher rate of rule denials on non-abstract problems than on symbolic problems.

When simple statements of $p$ and $q$ are viewed as observations about some world, $p \rightarrow q$ can be interpreted as a theory, or summarizing statement, about how the truth values of these observations are related.  This is, essentially, a model-theoretic viewpoint: an expression such as $p \rightarrow q$ is shorthand for how the truth values of $p$ and $q$ occur in the world. Taking this understanding of conditionals, the preference of our subjects to deny the conditional as a way of resolving contradiction can be interpreted as a preference to retain the truth value of "data" (the non-conditional sentences) and deny the particular interdependence that is asserted to hold between them.  This seems straightforwardly rational from an empiricist viewpoint: the "regularities" are nothing more than a way of summarizing the data. So, for a through-and-through empiricist, it is not even consistent to uphold a "law" in the face of recalcitrant data. Such a perspective puts a different light on the observation that people did not make the "easy" modus ponens inference from the expansion information combined with a modus tollens belief set: to have opted for this revision would have required changing the truth values of observational data. While doing so may be a plausible alternative when problems involve meaningless symbols, it may not seem rational alternative when working with information that is interpretable as observational data.

The idea that data enjoys a priority over regularities has been offered as a belief revision principle in other frameworks (Thagard, 1989; Harman, 1986) particularly when regularities are (merely) hypotheses under consideration to explain or systematize observed facts. There is a natural role, then, for induction mechanisms in specifying the process of belief revision, once the conditional "regularity" is chosen by the agent as suspect. We note that the classical AI belief revision community presents the belief revision problem as denying previously believed sentences, including conditionals. But replacing $p \longrightarrow q$ with $(p \,\&\, r) \rightarrow q$ or $(p \,\&\, {\sim}s) \longrightarrow q$ are equally good ways to deny $p \longrightarrow q$. In such a case, the conditional regularity can either be "patched" or demoted to the status of default rule ("Most of the time, $p \rightarrow q$, except when $r$ holds"). In our view, this method of denying a conditional as belief-revision choice seems to be preferable to merely lowering a degree of belief in the conditional, for the latter leaves the agent is no wiser about when to apply such a rule, only wiser that it should be less confident about the rule. This approach is being pursued in some classical approaches to belief revision (e.g., Ghose, Hadjinian, Sattar, You, and Goebel, 1993) and in explanation-based learning approaches to theory revision in the machine learning community, where the inability of a domain theory to explain some data causes changes to the domain theory rules (Ourston & Mooney, 1990; Richards & Mooney, 1995).

While some aspects of the belief revision process can be viewed as inductive processes searching for a better account of some data, we note that the such a perspective itself does provide principles for guiding such a process when there are alternative ways to reconcile a contradiction. Specifically, we don't *always* believe the data at the expense of a regularity or contingency that we currently believe holds in the world. As we noted in the introduction, there are intuitions opposite to those that would deny or change a regularity to accommodate data: Kyberg's (1983) belief framework includes a place for both measurement error and the knowledge that some types of observations are more prone to error than others. Thagard (1989) offers an explanatory coherence metric by which data can be discounted, if they cohere with hypotheses which themselves are poor accounts of a larger data set. Carlson & Dulany's (1988) model of reasoning with circumstantial evidence includes parameters for degrees of subjective belief in the evidence. So the broader questions for epistemic entrenchment might be to ask what kinds of data and what kinds of regularities are more differentially entrenched than others.

On our simple belief revision tasks, we found some baseline results that suggest a tendency to abandon the conditional. But it has long been recognized by researchers in both linguistics and human deduction that the if p then q form is used to express a broad range of different types of information, e.g., scientific laws, statistical relationships, causal relations, promises, and intentions ( "If it doesn't rain tomorrow, we will play golf"). More recent studies using both the selection paradigm described here as well as one in which subjects gave degrees-of-belief ratings to the initial belief sentences (Elio, 1996) have indicated that different types of knowledge expressed in this common syntactic form—causal relations distinguished by different enabling and disabling conditions, promises, and definitions—are differentially entrenched, when given new and contradictory evidence. For understanding the pragmatic principles of belief revision, these meta-knowledge distinctions may be important in formulating entrenchment preferences in the face of contradiction.

## On Multiple Models and Minimal Change

One clear result we obtained is that people retain uncertainty in their revised belief states—they did not use single models of the new information to chose among alternative interpretations of the initial information, or conversely, in the tasks we gave them (e.g., they did not follow Algorithm D). Further, they tended to select revisions that include more uncertainty than is logically defensible, opting for "*p* is uncertain and so is *q*" as often or more frequently than "*p* is true and *q* is false, or else *p* is false and *q* is true." It seems clear that people recognize that the former is less informative than the latter about possible combinations of *p* and *q*'s truth values, but our subjects chose it anyway. One way to view the results we obtained is to say that many of our subjects preferred revisions which were not minimal with respect to what was changed, but were instead minimal with respect to what they believed to hold true *without doubt* when both the initial and expansion information were considered jointly. It certainly seems more difficult to work with a "world" specification like {[*~p~q r s*] or [*p* *~q* *~r ~s*]} than it is with one that says "*q* is false and I'm not sure about anything else," even though (from a logical point of view) the former specification contains much more information than the latter.

What we learned from our initial investigations on minimal change problems may have less to do with the metrics of minimal change and more to do with issues of how people manipulate multiple models of the world. Rips' (1989) work on the knights-and-knaves problem also highlights the difficulty that people have in exploring and keeping track of multiple models. In that task, the supposition that one character is a liar defines one model, being a truth-teller defines another model, and each of these might in turn branch into other models. Even working such a problem out on paper presented difficulties for subjects, Rips reported. Yet in real life, we can certainly reason about vastly different hypothetical worlds that could be viewed as being equivalent to disjunctions of complex sentence sets. Unlike the arbitrary problems give to our subjects or even the knights and knaves problems, alternative hypothetical worlds about real-world topics may have some "explanatory glue" that holds together the particular contingencies, and no others, among the truth values of the independent beliefs. The question is whether for more real world situations, are people better able to retain and integrate the interdependencies among truth values in multiple models?

## Alternative Representations of Belief States

Representing a belief state as a set of sentences or even as a set of models is a simplification. We believe that number of important issues arise from this simple conceptualization and this study offers data on some of those issues. We noted alternative approaches to modeling belief states in the introduction, specifically those that use probabilistic information and degrees of belief. But there are two other perspectives that have long been considered from a philosophical viewpoint: the foundationalist view and the coherentist view. The foundationalist view (Swain, 1979; Alston, 1993; Moser, 1985, 1989) distinguishes between beliefs that are accepted without justification and those that depend on the prior acceptance of others. Such a distinction is used in truth-maintenance systems (e.g., Doyle, 1979; deKleer, 1986) for keeping track of dependencies among beliefs and to prefer the retraction of the latter ("assumptions") over the former ("premises") when contradictions are caused by new information. Pollock's (1987) defeasible reasoning theory defines a wider class of distinctions (e.g., "warrants" and "undercutters") and such distinctions can also be used to define normative foundationalist models of belief revision. The coherentist view (BonJour, 1985; Quine & Ullian, 1978; Harman, 1986) does not consider some beliefs as more fundamental than others, but rather emphasizes the extent to which an entire set of beliefs "coheres". One set of beliefs can be preferable to another if it has a higher coherence, however defined. Thagard's (1989) theory of explanatory coherence is an instance of this perspective and operational definitions of coherence can, in such a framework, be a means of implementing belief revision principles (Thagard, 1992). Pollock (1979) gives a whole range of epistemological theories that span the spectrum between foundationalist and coherentist.

It is widely believed (e.g., Harman, 1986; Gärdenfors, 1990b; Doyle, 1992; Nebel, 1992) that the original AGM account of belief revision, as well as model-based versions of it, are coherentist in nature. Harman (1986) and Gärdenfors go so far as to say that a foundationalist approach to belief revision (as advocated, e.g., by Doyle, 1979; Fuhrmann, 1991; Nebel 1991) is at odds with observed psychological behavior, particularly concerning people's ability to recall the initial justifications for their current beliefs. More marshaling of this and other experimental evidence (including the type we have reported in this article) could be a reasonable first step towards an experimentally-justified account of how human belief structures are organized; and with this is perhaps an account of how belief structures of non-human agents could best be constructed.

Finally, we note that it remains a difficult matter to examine "real beliefs" and their revision in the laboratory (as opposed to the task of choosing among sentences to be accepted as true); the paradigm of direct experimentation with some micro-world, which has been used to study theory development, is one direction that can prove fruitful (e.g., Ranney & Thagard, 1988). However, conceptualizing a belief state merely as a set of beliefs can still afford, we think, some insight into the pragmatic considerations people make in resolving contradiction.

## Future work

There are many issues raised in these investigations that warrant further study; we have touched upon some of them throughout our discussions. The possibility of bias against changing negated beliefs to non-negated ones, or in preferring revisions with non-negated sentences, needs systematic study. We used a selection paradigm throughout this study and it is important to establish whether similar results hold when subjects generate their new belief state. A more difficult issue is whether there are different patterns of belief revision depending on whether the belief set is one a person induces themselves or whether it is given to them. In the former case, one can speculate that a person has expended some cognitive effort to derive a belief, and a by-product of that effort may create the kind of coherentist structure that is more resistant to the abandonment of some beliefs in the face of contradictory information. This kind of perspective can be applied to an early study by Wason (1977) on self-contradiction. He found that subjects given the selection task were quite reluctant to change their conclusions about how to validate a rule, even when they were shown that such conclusions were contradicted by the facts of the task. Yet on the different sort of task, he found that subjects can recognize and correct invalid inferences about the form of a rule they are actively trying to identify from a data set, when the data set leads them to valid inferences that contradict the invalid ones they make. Whether recognizing contradiction depends on the demands a task makes of a reasoner might elucidate something about how premises are formulated and about how inferences are validated; in the belief revision scenarios we used in this study, the contradiction occurs not because of the reasoner's inferencing process, but because additional information about the world indicates that one of initially accepted premises must be suspect. The recognition and resolution of contradiction is important to general theories of human reasoning that employ deduction, induction, and belief revision. How general performance models of deductive and inductive reasoning can embrace belief revision decisions is an important open issue.

# References

Alchourrón, C., P. Gärdenfors, D. Makinson (1985). On the logic of theory change: Partial meet contraction and revision functions. Journal of Symbolic Logic, 50, 510-530.

Alston, W. (1993). The reliability of sense perception. Ithaca: Cornell University Press.

Ashton, R., & Ashton, A. (1990). Evidence-responsiveness in professional judgment: Effects of positive vs. negative evidence and presentation Mmode. Organizational Behavior and Human Decision Processes , 46, 1-19.

Bacchus, F., Grove, A., Halpern, J.Y., & Koller, D. (1992). From statistics to belief. In Proceedings of the Tenth National Conference on Artificial Intelligence, (pp. 602-608). Cambridge, MA: MIT Press.

BonJour, L. (1985). The structure of empirical knowledge. Cambridge: Harvard University Press.

Borgida, A. (1985). Language features for flexible handling of exceptions in information. Systems ACM Transactions on Database Systems, 10, 563-603.

Braine, M.D.S., & O'Brian, D. P. (1991). A theory of If: A lexical entry, reasoning program, and pragmatic principles. Psychological Review, 98, 182-203.

Carlson, R. A., & Dulany, D.E. (1988). Diagnostic reasoning with circumstantial evidence. Cognitive Psychology, 20, 463-492.

Cheeseman, P. (1988). Inquiry into computer understanding. Computational Intelligence, 4, 58-66.

Cheng, P. W., & Holyoak, K.J. (1989). On the natural selection of reasoning theories. Cognition, 33, 285-314.

Cheng, P.W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. (1993). Pragmatic versus syntactic approaches to training deductive reasoning. Cognitive Psychology, 18, 293-328.

Cheng, P.W., & Nisbett, R. E. (1993). Pragmatic constraints on causal deduction. In R.E. Nisbett (Ed.), Rules for reasoning. Hillsdale, NJ: Lawrence Erlbaum.

Cox, J. R., & Griggs, R. A. (1982). The effects of experience on performance in Wason's selection task. Memory & Cognition, 10, 496-502.

deKleer, J. (1986). An assumption-based TMS. Artificial Intelligence, 28, 127-162.

Dalal, M. (1988). Investigations into a theory of knowledge base revision: Preliminary report. Proceedings of the Seventh American Association for Artificial Intelligence, (pp. 475-479).

Doyle, J. (1979). A truth maintenance system. Artificial Intelligence, 12, 231-272.

Doyle, J. (1989). Constructional belief and rational representation. Computational Intelligence, 5, 1-11.

Doyle, J. (1992). Reason maintenance and belief revision: Foundations vs. coherence theories. In P. Gärdenfors (ed.) Belief revision, pp. 29-51. Cambridge: Cambridge University Press.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), Formal Representation of Human Judgment. NY: Holt Rinehart & Winston.

Einhorn, H., & Hogarth, R. (1978). Confidence in judgment: Persistence in the illusion of Validity. Psychological Review, 85, 395-416.

Elio, R. (1996). On the epistemic entrenchment of different types of knowledge expressed as conditionals. (Tech. Rep. TR96-16). Edmonton, Alberta: University of Alberta, Department of Computing Science.

Elio, R., & Pelletier, F. J. (1994). The effect of syntactic form on simple belief revisions and updates. In Proceedings of the 16th Annual Conference of the Cognitive Science Society. (pp. 260-265). Hillsdale, NJ: Lawrence Erlbaum.

Elmasri, R. & Navathe, S. (1994). Fundamentals of database systems, 2nd Edition. Redwood City, CA: Benjamin/Cummins.

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

Fagin, R., Ullman, J., & Vardi, M. (1986). Updating logical databases. Advances in Computing Research, 3, 1-18.

Foo, N.Y., & Rao, A.S. (1988). Belief revision is a microworld (Tech. Rep. No. 325). Sydney: University of Sidney, Basser Department of Computer Science.

Fuhrmann, A. (1991). Theory contraction through base contraction. Journal of Philosophical Logic. 20, 175-203.

Gärdenfors, P. (1984). Epistemic importance and minimal changes of belief. Australasian Journal of Philosophy, 62, 137-157.

Gärdenfors, P. (1988). Knowledge in flux: Modeling the dynamics of epistemic states. Cambridge, MA: MIT Press.

Gärdenfors, P. (1990a). Belief revision and nonmonotonic logic: Two sides of the same coin? In L. Aiello (ed.) Proceedings of the Ninth European Conference on Artificial Intelligence, Stockholm, pp. 768-773.

Gärdenfors, P. (1990b). The dynamics of belief systems: Foundations vs. coherence theories. Revue Internationale de Philosophie, 172, 24-46.

Gärdenfors, P., & Makinson, D. (1988). Revisions of knowledge systems using epistemic entrenchment. In Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge, (pp. 83-95). Los Altos, Calif.: Morgan Kaufmann.

Ghose, A.K., Hadjinian, P. O., Sattar, A., You, J., & Goebel, R. (1993). Iterated belief change: A preliminary report. In Proceedings of the Sixth Australian Conference on AI. Melbourne, pp. 39-44.

Halpern, J. Y. (1990). An analysis of first-order logics of probability. Artificial Intelligence, 46, 311-350.

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

Hoenkamp, E. (1988). An analysis of psychological experiments on non-monotonic reasoning. Proceedings of the Seventh Biennial Conference of the Canadian Society for the Computational Study of Intelligence. pp. 115-117.

Jeffrey, R.C. (1965). The logic of decision. New York: MacGraw Hill.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. Psychological Review, 99, 418-439.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). Deduction. Hillsdale, NJ: Lawrence Erlbaum.

Katsuno, H. & Mendelson, A. (1991). Propositional knowledge base revision and minimal change. Artificial Intelligence, 52, 263-294.

Koehler, J.J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. Organizational Behavior and Human Decision Processes, 56, 28-55.

Kyberg, H. E. Jr. (1983). Rational belief. Brain and behavioral sciences, 6, 231-273.

Kyberg, H.E. Jr. (1994). Believing on the basis of evidence. Computational Intelligence, 10, 3-20.

Lepper, M. R., Ross, L., & Lau, R.R. (1986). Persistence of inaccurate beliefs about the self: Perseverance effects in the classroom. Journal of Personality and Social Psychology, 50, 482-491.

Makinson, D., & Gärdenfors, P. (1991) Relations between the logic of theory change and nonmonotonic logic. In A. Fuhrmann & M. Morreau (eds.) The logic of theory change. Vol. 465 of Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Moser, P. (1985). Empirical justification. Dordrecht: D. Reidel.

Moser, P. (1989). Knowledge and evidence. Cambridge: Cambridge University Press.

Nebel, B. (1991). Belief revision and default reasoning: Syntax-based approaches. In Proceedings of the Second Conference on Knowledge Representation , (pp. 417-428) San Mateo, Calif.: Morgan Kaufmann.

Nebel, B. (1992). Syntax based approaches to belief revision. In P. Gärdenfors (ed.) Belief revision, pp. 52-88. Cambridge: Cambridge University Press.

Ourston, D., & Mooney, R. J. (1990). Changing the rules: A comprehensive approach to theory refinement. In Proceedings of the Eighth National Conference on Artificial Intelligence, (pp. 815-820). Cambridge, MA: MIT Press.

Pearl, J. (1988). Fusion, propagation, and structuring in belief networks. Artificial Intelligence, 29, 241-288.

Petty, R.E., Priester, J.R., & Wegener, D. T. (1994). Cognitive processes in attitude change. In R.S. Wyer & T.K. Srull (Eds.) Handbook of Social Cognition, Volume 2: Applications, (pp. 69-142). Hillsdale, NJ: Lawrence Erlbaum.

Pollock, J. L. (1979). A plethora of epistemological theories. In G. S. Pappas (Ed.), Justification and Knowledge: New Studies in Epistemology. pp. 93-113. Boston: D. Reidel.

Pollock, J. L. (1987). Defeasible reasoning. Cognitive Science, 11, 481-518.

Pollock, J. L. (1990). Nomic probabilities and the foundations of induction. Oxford: Oxford University Press.

Quine, W. & Ullian, J. (1978). The web of belief. NY: Random House.

Ranney, M. & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. In Proceedings of the Tenth Annual Conference of the Cognitive Science Society. (pp. 426-432). NJ: Lawrence Erlbaum.

Richards, B. L. & Mooney, R. J. (1995). Automated refinement of first-order horn-clause domain theories. Machine Learning, 19, 95-131,

Rips, L. J. (1983). Cognitive processes in propositional reasoning. Psychological Review, 90, 38-71.

Rips, L. J. (1989). The psychology of knights and knaves. Cognition, 31, 85-116.

Rips, L. J. (1994). The psychology of proof. Cambridge, MA: MIT Press.

Ross, L., & M. Lepper (1980). The perseverance of beliefs: Empirical and normative considerations. In R. Shweder (Ed.), Fallible Judgment in Behavioral Research. San Francisco: Jossey-Bass.

Russell, B. (1918). The philosophy of logical atomism. Reprinted in R. Marsh (ed.) Logic and knowledge. NY: Allen and Unwin, 1956.

Satoh, K. (1988). Nonmonotonic reasoning by minimal belief revision. In Proceedings of the International Conference on Fifth Generation Computer Systems, pp. 455-462. ICOT: Tokyo.

Shields, M.D., Solomon, I., & Waller, W. S. (1987). Effects of alternative sample space representations on the accuracy of auditors' uncertainty judgments. Accounting, Organizations, and Society, 12, 375-385.

Swain, M. (1979). Justification and the basis of belief. In G. S. Pappas (Ed.), Justification and knowledge: New studies inepistemology. Boston: D. Reidel.

Thagard, P. (1989). Explanatory coherence. Behavioral and Brain Sciences, 12, 435-502.

Thagard, P. (1992). Computing coherence. In R. Giere (ed.) Cognitive models of science. Minneapolis: University of Minnesota Press.

Wason, P. (1977). Self-contradictions. In P. Johnson-Laird & P. Wason (eds) Thinking: Readings in cognitive science. pp. 113-128. Cambridge: Cambridge University Press.

Weber, A. (1986). Updating propositional formulas. In Proceedings of the First Conference on Expert Database Systems, (pp. 487-500).

Willard, L., & Yuan, L. (1990). The revised Gärdenfors postulates and update semantics, In S. Abiteboul & P. Konellakis (eds) Proceedings of the International Conference on Database Theory, (pp. 409-421). Volume 470 of Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Wittgenstein, L. (1922). Tractatus Logico-Philosophicus. London: Routledge & Kegan Paul.

Yates, J.F. (1990). Judgment and decision making. Englewood Cliffs: Prentice Hall.

# Author Notes

This research was supported by Canadian NSERC research grants #A0089 to R. Elio and #OPG5525 to F. J. Pelletier. Portions of this work appeared in the Proceedings of the 16th Conference of the Cognitive Science Society, Atlanta, 1994. We thank Sarah Hoffman and Sioban Neary for their assistance in conducting these experiments, the Department of Psychology at the University of Alberta for the use of their subject pool, and Dr. Terry Taerem for discussions on statistical analyses. We are especially grateful to Aditya Ghose, Randy Goebel, Philip Johnson-Laird, Gil Harman, and an anonymous reviewer, whose comments on an earlier version of this manuscript greatly improved the presentation of this work.  Correspondence concerning this manuscript can be sent to Renée Elio, Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, T6G 2H1 or via email at ree@cs.ualberta.ca.

# Footnotes

[1] We note, however, that not all proponents of probabilistic frameworks concur that 'acceptance' is a required notion. Cheeseman (1988) and Doyle (1989), for example, argue that “acceptance” is really a mixture of two distinct components: the theory of degree of belief together with a theory of action. The latter theory uses degrees of belief plus a theory of utility to produce a notion of "deciding to act in a particular circumstance." Jeffrey (1965) also proposes a framework that avoids an acceptance-based account of belief.

[2] Most syntax-based approaches put into their definitions of belief revision that the set of all logical consequences is computed for the original belief state in order to determine the contradictions. But only changes to this original "base" belief set are considered in constructing the new belief state. One intuition behind this variety of belief revision can be that certain beliefs (the ones in the "base") are more fundamental than other beliefs, and any change in belief states should be made to the implicit beliefs first and only to the base if absolutely required. This view has relations to the foundationalist conception of belief states that we return to in our general discussion.

[3] Some works, e.g., Fagin, Ullman, and Vardi (1986), use the term "theory" to include both what we call a syntax-based approach and what we call a theory-based approach. When they want to distinguish the two, they call the latter a "closed theory."

[4] We aim to carefully distinguish our remarks about model-theoretic competence frameworks, as proposed by what we have been calling the classical AI belief revision community, from remarks concerning model-theoretic performance frameworks of human deduction, such as the mental-models theory. It is proper to talk of "models" in the context of either framework. Context will normally convey which framework we intend, but we use the term "formal AI models" or "mental models" when it is necessary.

[5] The explicit inclusion of ~r in S1 and ~n in S2 is, by some accounts, an extra inference step beyond what is necessary to incorporate ~w, since they could be considered as implicit beliefs rather than explicit beliefs; this could be accommodated simply by dropping r and n from S1 and S2, respectively.

[6] The actual problems used in these first experiments were really quantified versions of modus ponens and modus tollens. Our modus ponens problem type is more accurately paraphrased as: from For any x, if p holds of x, then q holds of x, and furthermore p holds of a we can infer q holds of a. Similar remarks can be made for our modus tollens.

[7] The reason is this. In a model approach, the initial belief state is the model [p is true, q is true]. When this is revised with ~q, thereby forcing the change from q's being true to q's being false in the model, we are left with the model [p is true, q is false]. Such a model has zero changes, other than the one forced by the expansion information; and in this model $\underline{p \rightarrow q}$ is false. In order to make this conditional be true, a change to the model that was not otherwise forced by the revision information would be required, to make p be false. (Similar remarks hold for the modus tollens case). Thus model theories of belief revision will deny the conditional in such problems.

[8] We used the term "knowledge" rather than "belief" in instructions to subjects, because we wanted them to accord full acceptance to them prior to considering how they might resolve subsequent contradiction. The use of "knowledge" here, as something that could subsequently change in truth value, is nonstandard from a philosophical perspective, although common in the AI community. Subsequent studies in which we called the initial belief set as "things believed to be true" have not impacted the type of results we report here.

[9] The loglinear model for this data is $\ln(F_{\underline{ijk}}) = \mu + \lambda\underline{r_i} + \lambda\underline{pres_j} + \lambda\underline{prob_k} + \lambda\underline{r_i pres_j} + \lambda\ \underline{r_i prob_k} + \lambda\ \underline{pres_j prob_k}$, where $F_{\underline{ijk}}$ is the observed frequency in the cell, $\lambda\underline{r_i}$ is the effect of the ith response alternative, $\lambda\underline{pres_j}$ is the effect of the jth presentation-form

category, $\lambda\mathrm{prob}_k$ is the effect of the *k*th problem type category, and the remaining terms are two-way interactions among these. The equivalent "logit" model, in which response is identified as the dependent variable, has terms for response, response by presentation mode, and response by problem type; it yields identical chi-square values. Loglinear and logit procedures from SPSS version 5.0 were used for these analyses. Simple chi-squares computed on several two-way frequency tables are consistent with the loglinear analyses and the conclusions presented in the text. The effect of symbol v. science-fiction presentation approached significance on both MP and on MT problems, when simple chi-squares were computed for separate two-dimensional frequency tables ($\chi^2$=5.65 and 4.87, p = .059 and .087; df=2 in both cases).

Table 1

Definitions of Initial Belief States and Revision Alternatives for Experiment 1's Problem Set

| Problem Type | | Revision Alternatives |
|---|---|---|
| Modus Ponens | | |
| Initial SS: | p—>q, p, q | 1. p—>q, ~p, ~q |
| Expansion: | ~q | 2. ~(p—>q), ~q ?p |
| | | 3. ~(p—>q) p ~q |
| Modus Tollens | | |
| Initial SS: | p —>q, ~p, ~q | 1. p—>q, p, q |
| Expansion: | p | 2. ~(p—>q), p, ?q |
| | | 3. ~(p—>q), p, ~q |

Note: SS means sentence set. Expansion means the expansion information. ? means uncertain.

Table 2

Percentage of subjects choosing each revision alternative, Experiment 1

| | Problem Type | | | | | |
|---|---|---|---|---|---|---|
| | Modus Ponens | | | Modus Tollens | | |
| Revision Alternative | Symbol | SciFi | Mean | Symbol | SciFi | Mean |
| 1. disbelieve initial non-conditional | .25 | .14 | .20 | .33 | .17 | .25 |
| 2. disbelieve conditional, uncertain about non-conditional | .38 | .29 | .34 | .38 | .54 | .46 |
| 3. disbelieve conditional | .37 | .58 | .48 | .28 | .29 | .29 |

Table 3

Percentage of subjects choosing each response alternatives, Experiment 2

| | Problem | |
|---|---|---|
| Revision Choice | Modus Ponens | Modus Tollens |
| 1. disbelieve non-conditional | .23 | .26 |
| 2. disbelieve conditional; non-conditional uncertain | .12 | .16 |
| 3. disbelieve conditional | .35 | .12 |
| 4. disbelieve both conditional and non-conditional | .14 | .02 |
| 5. both conditional and non-conditional uncertain | .16 | .44 |

Table 4

Templates for Experiment 3 problem types

| | | |
|---|---|---|
| Problem 1 | | |
| Initial Sentence Set | m & d —> g, m, d | |
| | [Therefore, g] | |
| Expansion | ~g | |
| Revision Alternatives | 1. | ~[m & d —> g], m, d |
| | 2. | m & d —> g, (~m & d) or (m & ~d) or (~m & ~d) |
| Problem 2 | | |
| Initial Sentence Set | c —> h, h —> m, c. | |
| | [Therefore, h and m] | |
| Expansion | ~h | |
| Revision Alternatives | 1. | h —> m, ~[c —> h], c, ?m |
| | 2. | h —> m, c —>h, ~c, ?m |
| | 3. | h —> m, ~[c —> h], c, m |
| | 4. | h —> m, c —> h, ~c, m |

Note: Bracketed consequences appeared in the initial sentence set for "consequences given" condition and were omitted in the "no consequences given" condition. All response choices included the expansion sentence as part of the revision description. See text for percentages of subjects choosing each option.

Table 5

Algorithms for Minimal Change

Algorithm D

```
D1 For each model of the expansion information do
        D1.1 For each model of the initial belief set do
                Find and save the differences.
        D1.2 From the set of differences, identify the smallest change. Put
                this smallest change and the expansion model responsible
                for it on the candidate stack.
D2.  From the candidate stack, chose as the new belief state the expansion model
        that is responsible for the smallest of all the minimal changes saved from D1.2.
        If there is more than one, use their disjunction
```

Algorithm W

```
W1      For each model of the belief set do
                W1.1   For each model of expansion do
                       Find and save the propositions that must change
                W1.2   Retain just the minimal set of propositions that must change for this
                       pairing of an belief set model and an expansion model
W2      Take the union of all proposition sets identified in 1.2 and remove them
        from the initial belief set
W3.     Identify the set of remaining KB propositions with known (certain) truth values.
                If this set is empty, then the new belief set is the expansion information.
                Otherwise, the new belief set is the conjunction of the old KB propositions
                with the expansion information
```

Table 5 continued next page

Table 5 continued

Algorithm B

```
B1.   For each model of the initial belief set   do
          B1.1 For each model of the expansion do
                  Find the differences and save them
          B1.2 From the set of differences, identify the minimal change and put the expansion model responsible for it on the candidate stack
B2.   Combine all models of expansion information on the candidate stack to determine the new belief state.
```

Algorithm S

```
S1    For each model of the initial belief set
          S1.1 For each model of the expansion, stack the differences between them.
S2.   From the set of differences, eliminate non-minimal changes
S3.   Combine all models of expansion information on the candidate stack to determine the new belief state.
```

Table 6

Problems and percentage of subjects choosing each revision alternative, Experiments 4 and 5

| | | | | Experiment | |
|---|---|---|---|---|---|
| Problem | | | Revision Alternative | 4 | 5 |
| 1 | Initial: | (pqrs ) or (~p~q~r~s) | 1. p ~q ~r ~s $_{D}$ | .06 | .07 |
| | Expansion: | p~r~s | 2. p ~r ~s ?q $_{B, S, W}$ | .94 | .58 |
| | | | 3. p q ~r ~s | | .05 |
| | | | 4. p ?q ?r ?s | | .30 |
| 2 | Initial: | (pqrs ) or (~p~q~r~s) | 1. p ~q ~r ~s $_{D}$ | .11 | .07 |
| | Expansion: | (~p~qrs) or (p~q~r~s) | 2. (~p~qrs) or (p~q~r~s) $_{B, S, W}$ | .89 | .21 |
| | | | 3. ~p ~q r s | | .35 |
| | | | 4. ~q ?p ?r ?s | | .37 |
| 3 | Initial: | pq~s | 1. ~p ~q r ~s $_{D, S, W}$ | .22 | .20 |
| | Expansion: | (~p~q) & [(r~s) or (~rs)] | 2. (~p~q) & [(r~s) or (~rs)] $_{B}$ | .78 | .43 |
| | | | 3. ~p ~q ~r s | | .0 |
| | | | 4. ~p ~q ?r ?s | | .37 |
| 4 | Initial: | p q | 1. p or q, not both $_{D, B, S}$ | .12 | .07 |
| | Expansion: | ~p or ~q or (~p~q) | 2. ~p or ~q or (~p~q) $_{W}$ | .88 | .30 |
| | | | 3. ~p ~q | | .12 |
| | | | 4. ?p ?q | | .51 |

(Table 6 continued next page)

Table 6 continued

| Problem | | | Revision Alternative | Experiment | |
|---|---|---|---|---|---|
| | | | | 4 | 5 |
| 5 | Initial: | pqr | 1. ~p q r $_{D, B}$ | .10 | .21 |
| | Expansion: | (~pqr) or (p~q~r) | 2. (~p q r) or (p ~q ~r) $_{S, W}$ | .90 | .26 |
| | | | 3. p ~q ~r | | .07 |
| | | | 4. ?p ?q ?r | | .46 |
| 6 | Initial: | ~p~q~r | 1. p ~q ~r $_{D, B}$ | | .07 |
| | Expansion: | (~pqr) or (p~q~r) | 2. (~p q r) or (p ~q ~r) $_{S,W}$ | | .30 |
| | | | 3. ~p q r | | .23 |
| | | | 4. ?p ?q ?r | | .40 |

Note: Initial means initial sentence set. Expansion means expansion information.

# Appendix A

## Clauses used for Science-Fiction Stimuli-, Experiment 1

Subjects received one of the three possible science-fiction versions of the modus ponens and modus tollens rules, given below. Each version was used equally often across subjects.

### Modus Ponens Rules

If a Partiplod hibernates during the day, then it is a meat eater.

If a cave has a Pheek in it, then that cave has underground water.

If a ping burrows underground, then it has a hard protective shell.

### Modus Tollens Rules

If Gargons live on the planet's moon, then Gargons favor interplanetary cooperation.

If an ancient ruin has a force field surrounding it, then it is inhabited by aliens called Pylons.

If a Gael has cambrian ears (sensitive to high-frequency sounds), then that Gael also has tentacles.

## Appendix B
## Phrasing of Problems in the Symbolic Condition for Experiments 4 and 5

| Problem | Initial Belief Set | Expansion |
|---|---|---|
| 1 | Either A, B, C, and D are all true, or none of them are true. | A is true. C is true. D is false. |
| 2 | Either A, B, C, and D are all true, or none of them are true. | B is false. Exactly one of these is true, but no one knows for sure which one:<br>• A is true, and C and D are both false.<br>• A is false, and C and D are both true. |
| 3 | A is true. B is true. D is true. | A is false. B is false. Either C is true or D is true, but not both of them. |
| 4 | A is true. B is true. | At least one of A and B is false, and possibly both of them are. |
| 5 | A is true. B is true. C is true. | Either A is false and Band C are both true, or A is true and B and C are both false. No one knows for sure which it is. |
| 6 | A is false. B is false. C is false. | Either A is true and B and C are both false, or A is false and B and C are both true. No one knows for sure which it is. |