First-order belief revision

Samir Chopra and Eric Martin

School of Computer Science and Engineering
University of New South Wales
Sydney, NSW 2052, Australia
schopra,emartin@cse.unsw.edu.au

Abstract. We present a model for first-order belief revision that is characterized by an underlying relevance-like relation and a background proof system. The model is extremely general in order to allow for a wide variety in these characterizing parameters. It allows some weakenings of beliefs which were initially implicit to become explicit and survive the revision process. The effects of revision are localized to the part of the theory that is influenced by the the new information. Iterated revision in this model is handled trivially since the revision operator is constructive by definition. The usage of deductively limited proof systems permits an inconsistency tolerant model. The notion of a part of a theory capable of being influenced by new information (designed to accomodate the specific character of first-order languages) is shown to satisfy some intuitive and desirable properties. We show that for particular parametrizations, standard revision schemes can be embedded into our paradigm.

1 Introduction

Why does the need for a "genuine" theory of first-order belief revision arise? The AGM [2] postulates demarcate a class of belief revision operators but are remarkably abstract and high-level when it comes to the language that is assumed to underly the corresponding revision operations. The AGM postulates deal with the set-theoretic properties of the revision process (in demarcating a class of minimal change operators) and not with the non-Boolean aspects of the background logic. The theories that are the target of the revisions could easily be propositional, modal S_5, first-order, and so on. The postulates make no specific reference to the syntactic structure of the theories other than their propositional components. In designing a theory for "genuine" first-order revision, then, we demand that *the revision operation and background logic address the particular constraints created by the presence of quantifiers in the underlying language*. This is a basic technical requirement. What about rationality, plausibility and computational requirements? Models for belief revision should satisfy the following desiderata:

- Plausible representations of epistemic states of agents must be finite.
- Since agents typically only reason with a limited portion of their entire epistemic state, some notion of context sensitivity or relevance guides them in their choice of premises for any inference operation.
- To be genuinely capable of implementation in automated reasoning systems the revision procedures must be tractable.

- A distinction must be made between the explicit and implicit beliefs of the agent.
- Agents are often inconsistent, knowingly or unknowingly. Without making claims about the rationality of such agents, it is uncontroversial that any plausible model must be inconsistency tolerant without lapsing into incoherence or triviality.

Our model is compatible with underlying logics which can be complete, incomplete, or even unsound (because, for instance, they incorporate inductive rules). This allows us to capture the intuitions such as the agent not being fully logically omniscient, due to resource limitations. We present the receipt of new information by the agent as an awareness triggering process. Firstly, the incoming data determines which members of the belief set will be subject to revision. Secondly, the agent calculates the consequences (according to the underlying logic) of the new information in conjunction with those beliefs. Finally, any detected inconsistencies are removed by the selective retainment of some of the calculated consequences. As a result, our framework strikes a happy balance between models of belief revision that blur the distinction between implicit and explicit beliefs, and those that concentrate on just the latter. At any given stage, the epistemic state of the agent consists of beliefs that are explicitly believed by the agent, and some beliefs that were initially implicit but have been made explicit by the arrival of new information. Most significantly, we formalize a class of logics for inquiry: the use of relevance-like concepts—technically defined as triggering functions below—corresponds to the selection of premises by the agent and the associated rules manipulate those premises. The revision scheme is intended to be extremely general; given a particular choice of a triggering function and an inference scheme, a revision operator is generated. Iterated revision in this model is handled trivially since the revision operator is constructive by definition.

The outline of the paper is as follows: we present the notion of triggering functions, motivate it with examples and discussion, describe the framework of the revision scheme and then move on to a brief discussion of some basic formal properties. In particular, we show that full meet AGM revision for bases and closed theories (with minor caveats) is obtained naturally.

Notation: We fix a *vocabulary* S, that is a set of function symbols (possibly including constants) and predicate symbols, and we denote by $\mathfrak{L}$ the set of all first-order, equality free S-formulas. The set of all atomic members of $\mathfrak{L}$ is written as $\mathfrak{L}_{\text{at}}$. For all $\varphi \in \mathfrak{L}$, $\text{At}(\varphi)$ denotes the set of atomic formulas occurring in φ. We denote by $\bot$ an extra formula, not belonging to $\mathfrak{L}$, assumed to be false in every S-structure. Hence the notation $D \models \bot$, where D is a subset of $\mathfrak{L}$, will mean naturally that D is inconsistent. Given $T \subseteq \mathfrak{L}$, $\text{Mod}(T)$ denotes the class of all S-structures that are models of (the universal closure of each member of) T. Given S-structure $\mathfrak{M}$, $|\mathfrak{M}|$ denotes the domain of $\mathfrak{M}$. We use $X \setminus Y$ to represent the difference of two sets X, Y.

2 Triggering functions

In this section, we define a range of functions intended to capture differing intuitions about relevance or relatedness between formulas. Often, relevance relations may be felt to be too fine grained, often they might be judged to be too coarse. Our objective is to provide a definition that captures the entire range and to then investigate what sorts of revision schemes will be captured by particular instances of these functions.

Definition 1 *A* triggering function *is a total mapping $\tau : \mathcal{L} \to 2^{\mathcal{L}}$ such that:*

1. *for all $\varphi \in \mathcal{L}$, At$(\varphi) \subseteq \tau(\varphi)$;*
2. *for all $\varphi, \psi \in \mathcal{L}$, if At$(\varphi) \subseteq$ At(ψ) then $\tau(\varphi) \subseteq \tau(\psi)$.*

As a consequence of clause *2.* in the definition above, note that $\tau(\varphi) = \tau(\neg\varphi)$ for all $\varphi \in \mathcal{L}$. If τ is a triggering function and φ a formula, $\tau(\varphi)$ intuitively provides a set of "building blocks" (atomic formulas) including all of those from which φ can be constructed. Any formula ψ which uses at least one of these "building blocks" in its construction will then be considered relevant to φ. The coarsest triggering function is constant: it maps every member of $\mathcal{L}$ to $\mathcal{L}_{\mathrm{at}}$. Choosing this triggering function will effectively mean working with no notion of relatedness between formulas. The finest triggering function maps every $\varphi \in \mathcal{L}$ to At(φ). As an intermediate case, consider the triggering function that maps every $\varphi \in \mathcal{L}$ to the set of all $\psi \in \mathcal{L}_{\mathrm{at}}$ such that some symbol in S occurs in both φ and ψ. We call this particular function the *crude* triggering function. The intuition behind it is simple: if formula φ triggers formula ψ, then ψ refers to some of the properties, functions or individuals referred to by φ, and vice versa. From now on, τ denotes an arbitrary triggering function.

Definition 2 *Let $\varphi, \psi \in \mathcal{L}$ be given. We say that φ τ*-triggers *ψ iff some atomic formula occurring in ψ is an instance of a member of $\tau(\varphi)$. We denote by Trig$_\tau(\varphi)$ the set of members of $\mathcal{L}$ that φ τ-triggers.*

Observe that for all $\varphi \in \mathcal{L}$, φ belongs to Trig$_\tau(\varphi)$.

Property 1 *Suppose that τ is the crude triggering function. Let $\varphi, \psi \in \mathcal{L}$ be given. The following conditions are equivalent.*

1. *φ τ-triggers ψ;*
2. *ψ τ-triggers φ;*
3. *some symbol in S occurs in both φ and ψ.*

Example 1 *Suppose that S consists of a constant c, a unary function symbol f and a binary function symbol g, a unary predicate symbol P and two binary predicate symbols Q and R. Also suppose that τ is the crude triggering function. Set:*

- $\varphi_1 = \exists x \forall y((P(x) \vee R(c, g(x, y))) \to P(z))$;
- $\varphi_2 = \exists x R(x, g(c, c)) \vee \forall x R(x, c)$;
- $\psi = \exists x(P(x) \vee \forall y Q(x, y)) \to Q(z, f(z))$.

Then:

1. *At$(\varphi_1) = \{P(x), R(c, g(x, y)), P(z)\}$, At$(\varphi_2) = \{R(x, g(c, c)), R(x, c)\}$, and At$(\psi) = \{P(x), Q(x, y), Q(z, f(z))\}$.*
2. *$\psi \in$ Trig$_\tau(\varphi_1)$ since P occurs in φ_1.*
3. *$\psi \notin$ Trig$_\tau(\varphi_2)$ since neither of P, Q and f occurs in φ_2.*

In the case of first-order languages, the crude triggering function makes a second-order statement since it involves not just individuals but also predicates. In the example above, to say that ψ is τ-triggered by φ_1 is to assert the existence of a property (P) that is common to the individuals of the domain. In some cases, the use of the crude triggering function might not mesh with all our intuitions about relevance. For instance, consider the following: $P(c)$ is τ-triggered by $\forall x Q(c, x)$ whereas $\forall x P(x)$ is not τ-triggered by $\forall x Q(c, x)$. This judgement might be disputable[1] , since $\forall x P(x)$ implies $P(c)$. Conversely, the crude triggering function might be too coarse. For instance, $P(f(d))$ (where d is a constant) and $\forall x Q(f(c), x)$ are τ-triggered by each other because of the function symbol f appearing in both while clearly, $f(d)$ and $f(c)$ might refer to distinct objects (of course, a contrary intuition could be that the two formulas *are* relevant since they both tell us something about the function f and its arguments). In the following discussion, we will often make use of the crude triggering function, despite its shortcomings since it is easily calculated and does not add excessively to the technical complexity of our framework[2] . We remark that providing a formal definition of a relevance relation that does justice to *all* intuitions is a task that still awaits completion.

Proposition 1 below shows that our general notion of relevance satisfies the fundamental requirement that if two sets of sentences are irrelevant to each other, then they cannot be in conflict. An informal argument for the proposition is as follows: let triggering function τ and formulas φ, ψ be such that φ does not τ-trigger ψ. Then none of the building blocks provided by τ on the basis of φ is used in the construction of ψ. By condition *1.* in Definition 1, this guarantees that given any structure $\mathfrak{M}$, the interpretations of φ, ψ in $\mathfrak{M}$ are unrelated and independent.

Proposition 1 *Let $\{T_i \mid i \in I\}$ be a set of consistent subsets of $\mathcal{L}$ such that for all distinct $i, j \in I$, no formula in T_i τ-triggers a formula in T_j. Then $\bigcup_{i \in I} T_i$ is consistent.*

Proof. Say that S-structure $\mathfrak{M}$ is standard if the following holds.

1. Let $n, m \in \mathbb{N}$, n-ary function symbol f in S, m-ary function symbol g in S, and $a_1 \dots a_n, b_1 \dots b_m \in |\mathfrak{M}|$ be given. Then $(f(a_1 \dots a_n))^{\mathfrak{M}} = (g(b_1 \dots b_m))^{\mathfrak{M}}$ iff $f = g$ (which implies that $m = n$) and for all $1 \le i \le n$, $a_i = b_i$.
2. Define the base of $\mathfrak{M}$ as the set of all members a of $|\mathfrak{M}|$ such that for every $n \in \mathbb{N}$, n-ary function symbol f in S, and $a_1 \dots a_n \in |\mathfrak{M}|$, $(f(a_1 \dots a_n))^{\mathfrak{M}} \neq a$. Then the base of $\mathfrak{M}$ is infinite.

Since $\mathcal{L}$ is equality free, it is easy to verify that every consistent subset of $\mathcal{L}$ has a standard model. For all $i \in I$, choose a standard model $\mathfrak{M}_i$ of T_i. Say that two S-structures $\mathfrak{M}$ and $\mathfrak{N}$ are similar just in case $\mathfrak{M}$ and $\mathfrak{N}$ have the same domain, have the same base, and are such that for all $n \in \mathbb{N}$, n-ary function symbol f in S, and $a_1 \dots a_n$ in the common domain of $\mathfrak{M}$ and $\mathfrak{N}$, $(f(a_1 \dots a_n))^{\mathfrak{M}}$ and $(f(a_1 \dots a_n))^{\mathfrak{N}}$ are equal.

[1] This would require alternative triggering functions so as to make both $P(c)$ and $\forall x P(x)$ related to $\forall x Q(c, x)$). Consider for instance the following: for all $\varphi \in \mathcal{L}$, define $\tau(\varphi)$ as the set of all formulas of the form $R(t_1 \dots t_n)$ where R is an n-ary predicate symbol in S, and all the t_i's are closed terms occuring in at least one instance of a member of $At(\varphi)$.

[2] It has been pointed out to us that the crude triggering function is invariant under Skolemisation—a fundamental requirement.

Without loss of generality we can assume that for all $i, j \in I$, $\mathfrak{M}_i$ and $\mathfrak{M}_j$ are similar. Let $i \in I$ be given. Define X_i to be the set of all expressions of form $P(a_1 \ldots a_n)$, where n is a nonnul integer, P is an n-ary predicate symbol in S, and there exists S-terms $t_1 \ldots t_n$, pairwise distinct variables $x_1 \ldots x_m$ that exhaust the variables occurring in at least one of $t_1 \ldots t_n$, and $b_1 \ldots b_m \in |\mathfrak{M}_i|$ such that $P(t_1 \ldots t_n) \in \bigcup\{\mathrm{At}(\varphi) \mid \varphi \in T_i\}$ and $(t_1[b_1/x_1 \ldots b_m/x_m])^{\mathfrak{M}_i} = a_1, \ldots, (t_n[b_1/x_1 \ldots b_m/x_m])^{\mathfrak{M}_i} = a_n$. Since for all distinct $i, j \in I$, no formula in T_i τ-triggers some formula in T_j, we infer easily that the following holds.

1. For all distinct $i, j \in I$, $X_i \cap X_j = \emptyset$.
2. Let S-structure $\mathfrak{M}$ be such that $\mathfrak{M}$ is similar to $\mathfrak{M}_i$ and for all $\varphi \in X_i$, $\mathfrak{M} \models \varphi$ iff $\mathfrak{M}_i \models \varphi$. Then $\mathfrak{M} \models T_i$.

By the first point, there exists an S-structure $\mathfrak{M}$ similar to the $\mathfrak{M}_i$'s and such that for all $i \in I$ and $\varphi \in X_i$, $\mathfrak{M} \models \varphi$ iff $\mathfrak{M}_i \models \varphi$. By the second point, $\mathfrak{M} \models \bigcup_{i \in T} T_i$. Hence $\bigcup_{i \in T} T_i$ is consistent. $\Box$

As a corollary of the previous proposition, the deduction of a formula can be localized to that part of the theory that is relevant to it.

Corollary 1 *Let consistent $T \subseteq \mathfrak{L}$ and $\varphi \in \mathfrak{L}$ be such that $T \models \varphi$. Let T' be the set of all formulas ψ in T such that either φ τ-triggers ψ or ψ τ-triggers φ. Then $T' \models \varphi$.*

Triggering functions enable us to deal with inconsistent theories in an interesting and natural way since we can localize inconsistencies to the set of all formulas triggered by a formula (a current belief) in some initial set of beliefs.

Definition 3 *Let $T \subseteq \mathfrak{L}$ be given. We say that T is* locally τ-consistent *just in case for all $\varphi \in T$, $Trig_\tau(\varphi) \cap T$ is consistent.*

Assessing the consistency of an agent relative to a given formula φ should enforce the consideration of all formulas related to φ. For some theories T, triggering functions serve to subdivide the theory into disjoint sets of unrelated sentences. This can be viewed as the division of the theory into beliefs about unrelated subject matters as expressed by the definition below. The subsequent corollary of Proposition 1 shows that local consistency guarantees consistency.

Definition 4 *A subset T of $\mathfrak{L}$ is said to be* τ-partitioned *iff $\{Trig_\tau(\varphi) \cap T \mid \varphi \in T\}$ is a partition of T.*

Corollary 2 *A τ-partitioned subset of T is consistent iff it is locally τ-consistent.*

3 Revision

We now present our revision scheme. In what follows, we shall refer to revision models that use closed and consistent theories to represent belief sets as AGM-revision schemes (e.g. [2]); those using open sets of sentences will be referred to as base revision schemes (e.g. [10, 7]). In both schemes, the agent is assumed to have the entire deductive machinery of classical logic at its disposal. In particular, the agent is assumed to have the ability of determining the inconsistency of a set of sentences. In contrast, we would like to study the revision strategies of agents with limited resources for deduction *i.e.*, they might not possess a complete set of inference rules. This has the consequence that given a possibly inconsistent finite set of sentences, the agent might not be able to infer a contradiction. In both revision schemes, whether a formula is dropped from an agent's belief set or not depends on the formula being implicated in a contradiction. The notion of proof is implicit whereas in our model it is explicit since no standard deductive machinery is assumed. Therefore, in our model a formula's retainment will be decided on the basis of its presence in a particular deduction of a contradiction. Moreover, making the proofs explicit enables us to address the issue of whether the logical consequences of an implicated formula should be retained or not. In contrast, note that in AGM and base revision schemes, this is not possible since on one extreme, belief sets must be closed under logical consequence before and after revision and on the other, in base revision, the issue is simply ignored. Another concern sidestepped in both schemes is the issue of relevance. We would like, however, to only revise that part of the belief set that is relevant to the new information. The section above has provided us with the required formal tools.

Moving on to technical details, our revision scheme will be characterized by an associated triggering function and background logic which is described in terms of proof trees. The triggering function formalizes the relevance indicator used by the agent. The proof trees are a natural and convenient representation of a set of inference rules.

Definition 5 *A* proof tree *is a finite labelled tree τ with labels drawn from $\mathfrak{L} \cup \{\#, \bot\}$, where $\#$ denotes a symbol not in $\mathfrak{L} \cup \{\bot\}$, such that:*

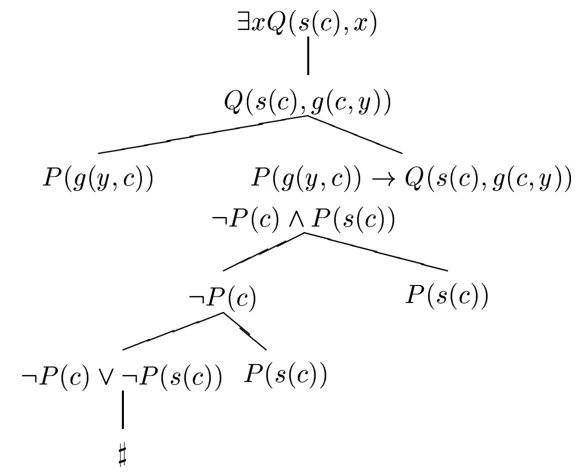
1. *$\#$ labels only some (maybe none) of the leaves in τ;*
2. *$\bot$ possibly labels the root of τ, but no other node in τ.*

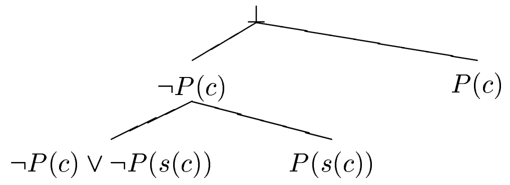
A proof tree is e.c. *(for* explicitly contradictory*) just in case its root is labelled with $\bot$; otherwise it is* n.e.c. *(for* non explicitly contradictory*).*

A proof tree τ is valid *iff for all nodes N in τ, if $\varphi \in \mathfrak{L} \cup \{\bot\}$ labels N and D is the set of labels in $\mathfrak{L}$ of the internal nodes of the subtree of τ whose root is N, then $D \models \varphi$.*

A formula labeling a leaf should be viewed as an axiom *i.e.*, a member of the original belief set or the new information itself. Proofs sometimes involve tautologies (consequences of the empty set) which might not explicitly occur in the belief set. The occurrence of $\#$ represents the empty set from which these tautologies are derived.

Example 2 *Suppose that S is defined as in the previous examples. Here are three examples of proof trees. The first two ones are n.e.c., the last one is e.c. Only the first and the third one are valid. The second is not, since the node whose child is labeled with $\#$ is labeled with a non-tautological formula.*





Note that if a proof tree whose root is labelled with $\perp$ is valid, then the set of formulas which label its leaves is inconsistent. But the converse does not necessarily hold. Given $n \in \mathbb{N}$, n.e.c. proof trees $\tau_1 \ldots \tau_n$, and $\varphi \in \mathcal{L}$, we denote by $\langle \tau_1 \ldots \tau_n, \varphi \rangle$ the proof tree whose root is labelled with φ and has $\tau_1 \ldots \tau_n$ as children (the condition includes the case $n = 0$). We use the same notation if φ denotes $\perp$ instead of a member of $\mathcal{L}$. We identify a tree consisting of a single node with its label.

Definition 6 *A* derivation frame *is a set $\mathcal{D}$ of proof trees with the following properties.*

1. *$\mathcal{D}$ is closed under subtrees.*
2. *$\mathcal{D}$ is closed under the substitution of leaves in members of $\mathcal{D}$ by members of $\mathcal{D}$ whose roots have the same label as the label of the leaf being substituted.*

A derivation frame is valid *iff all its members are valid.*

Clearly, for all sets X of proof trees, there exists a smallest derivation frame which contains X, and which is valid iff all the members of X are valid. From now on, $\mathcal{D}$ denotes a derivation frame. In practice, $\mathcal{D}$ represents all the proofs that can be built from $\mathcal{L}$ by repeated applications of inference rules taken from a given set I; in this case, we say that $\mathcal{D}$ is *generated* by the rules in I. Suppose for instance that the set of rules I contains modus ponens. If $\mathcal{D}$ is the derivation frame generated by I, and if $\varphi, \psi \in \mathcal{L}$ and $\sigma, \tau \in \mathcal{D}$ are such that the root of σ is labelled with φ and the root of τ is labelled with $\varphi \to \psi$, then $\langle \sigma, \tau, \psi \rangle$ is a member of $\mathcal{D}$. The set I could be one of the many sets of inference rules shown to be complete. Or it could be a more restricted one, for instance because it accepts only "relevant" inference rules, for some meaning of "relevant." It could even contain inference rules which are not necessarily valid (*i.e.*, such that the consequent is not necessarily a logical consequence of the premises), which would correspond to some kind of nonmonotonic, or inductive logic, etc. In that case, the derivation frame corresponding to I would fail to be valid according to the definition above.

The special formula $\perp$ represents an explicit contradiction. It could be identified with (*i.e.*, replace as label of the root of a proof tree) a given contradiction, or any contradiction, etc. It could also be introduced in light of a rule of the kind: from $\{\varphi, \neg\varphi\}$ derive $\perp$, for any $\varphi \in \mathcal{L}$. In that case, $\perp$ would acknowledge an explicit, blatant contradiction, and we would have that a proof tree in the corresponding derivation frame is e.c. just in case it is of the form $\langle \sigma, \tau, \perp \rangle$, where σ and τ are n.e.c. proof trees, the root of one being labelled with φ and the root of the other with $\neg\varphi$, for some $\varphi \in \mathcal{L}$. In case a set I of inference rules contains a rule with no premises (for example, from the empty set derive any tautology), then the derivation tree corresponding to one application of this rule would have the form $\langle \#, \varphi \rangle$. This should be distinguished from the proof tree $\langle \varphi \rangle$, whose intended meaning is: φ is an axiom *i.e.*, a member of the theory from which proofs are conducted. More precisely, we have the following:

Definition 7 *Let $T \subseteq \mathcal{L}$ be given. A* $\mathcal{D}$-proof from T *is a proof tree in $\mathcal{D}$ all of whose leaves' labels are drawn from $T \cup \{\#\}$. If $\varphi \in \mathcal{L} \cup \{\perp\}$ labels the root of a $\mathcal{D}$-proof σ from T, then we say that σ is a* $\mathcal{D}$-proof of φ from T.

We now explain the revision strategy of our model. When a new piece of information φ is received by the agent, the triggering function is used to determine the set of formulas in the theory relevant to it. Only this portion of the theory extended with φ, say Z, will be subject to revision. The remainder is unaffected. The agent then employs the inference rules at its disposal to calculate the consequences of Z. When explicit contradictions are detected, if any, then all formulas in Z used in the derivation are dropped. All other formulas in Z are retained. In a significant departure from standard revision strategies, some consequences of Z (possibly involving dropped formulas) will be retained as well. The criterion for retaining these formulae is the same: they must not be implicated in the derivation of a contradiction. This makes possible the retention of useful consequences (which were merely implicit before revision) of formulas that are found to be inconsistent with the new information. All retained formulas are then merged (explicitly) with the unaffected beliefs to obtain the agent's new belief set. This process is *iterable* since a precise construction of the new belief set is provided.

Definition 8 *The* revision function determined by τ and $\mathcal{D}$ *is the total function $\oplus^{\tau}_{\mathcal{D}}$: $2^{\mathcal{L}} \times \mathcal{L} \to 2^{\mathcal{L}}$ defined as follows. Let $T \subseteq \mathcal{L}$ and $\varphi \in \mathcal{L}$ be given. Then $\oplus^{\tau}_{\mathcal{D}}(T, \varphi)$ (usually written $T \oplus^{\tau}_{\mathcal{D}} \varphi$) is equal to $X \cup Y \cup \{\varphi\}$, where:*

1. *X is the set of all members of T that are not τ-triggered by φ;*
2. *Y is the set of all members of $\mathcal{L}$ that occur in at least one n.e.c. $\mathcal{D}$-proof from $(T \cup \{\varphi\}) \cap Trig_{\tau}(\varphi)$, but in no e.c. $\mathcal{D}$-proof from $(T \cup \{\varphi\}) \cap Trig_{\tau}(\varphi)$.*

Instead of $\oplus^{\tau}_{\mathcal{D}}$, we use the notation $\oplus_{\mathcal{D}}$ when we (tacitly) assume that τ is the coarsest triggering function. Clearly, our definitions can be immediately generalized from revision by single formulas to revision by (possibly infinite) sets of formulas. This would enable us to use the revision function in incremental or batch fashion. That is, for an initial theory T and a sequence $\varphi_0 \dots \varphi_n$ of inputs, we could either compute $T \oplus^{\tau}_{\mathcal{D}} \varphi_0 \oplus^{\tau}_{\mathcal{D}} \varphi_1 \oplus^{\tau}_{\mathcal{D}} \dots \oplus^{\tau}_{\mathcal{D}} \varphi_n$ or $T \oplus^{\tau}_{\mathcal{D}} \{\varphi_0 \dots \varphi_n\}$. We stick to the simpler definition since it conveys the key ideas of our framework without loss of generality. The following example illustrates our revision scheme.

Example 3 *Suppose that τ is the crude triggering function and that S consists of constants a, b, c and unary predicates P, Q. Let $\mathcal{D}$ be the derivation frame generated by the following three inference rules:*

$$\frac{\varphi, \varphi \to \psi}{\psi} \quad \textit{(MP)}$$

$$\frac{\forall x \varphi}{\varphi[t/x]} \quad \textit{where } t \textit{ is a closed } S\textit{-term (UI)}$$

$$\frac{\varphi, \neg\varphi}{\perp} \quad \textit{(BC)}$$

Consider $T = \{\forall P(x) \to Q(x), \neg Q(a), \neg Q(b)\}$ and $\varphi = P(a)$. Then $Trig_{\tau}(\varphi) \cap T = \{\forall P(x) \to Q(x), \neg Q(a)\}$. In the proof trees shown below, the first represents the detection of a contradiction via the application of MP, UI and BC. The second and third represents the application of the rule UI.

$$\dfrac{\dfrac{\dfrac{\forall x(P(x) \to Q(x))}{P(a) \to Q(a)} \qquad P(a)}{Q(a)} \qquad \neg Q(a)}{\perp}$$

$$\dfrac{\forall x(P(x) \to Q(x))}{P(b) \to Q(b)} \qquad\qquad \dfrac{\forall x(P(x) \to Q(x))}{P(c) \to Q(c)}$$

The set of formulas that occur in at least one n.e.c. $\mathcal{D}$-proof from $(T \cup \{\varphi\}) \cap Trig_{\tau}(\varphi)$ is $\{P(a), Q(a), \neg Q(a), P(a) \to Q(a), \forall x(P(x) \to Q(x)\}$. The set of formulas that occur in no e.c. $\mathcal{D}$-proof from $(T \cup \{\varphi\}) \cap Trig_{\tau}(\varphi)$ is $\{P(b) \to Q(b), P(c) \to Q(c)\}$. We conclude that $T \oplus^{\tau}_{\mathcal{D}} \varphi = \{\neg Q(b), P(b) \to Q(b), P(c) \to Q(c), P(a)\}$.

Note the significant features of the revision process: only a part of the theory is considered a candidate for revision, weakenings of quantified formulas are retained and only formulas implicated in explicit contradictions are dropped.

The following shows that if the derivation frame is valid then the revision process is sound.

Property 2 *Suppose that $\mathcal{D}$ is valid. For all $T \subseteq \mathcal{L}$ and $\varphi \in \mathcal{L}$, $T \cup \{\varphi\} \models T \oplus^{\tau}_{\mathcal{D}} \varphi$.*

The following shows that given a derivation frame that is capable of detecting inconsistencies, a consistent input and with no relevance relation to guide the revision process, the result of revisions is always a consistent theory. This is a baseline property for any plausible belief revision scheme.

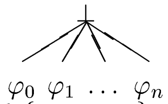
Proposition 2 *Suppose that $\mathcal{D}$ is valid and for all inconsistent $D \subseteq \mathcal{L}$, there exists a $\mathcal{D}$-proof of $\bot$ from D. Then for all $T \subseteq \mathcal{L}$ and satisfiable $\varphi \in \mathcal{L}$, $T \oplus_{\mathcal{D}} \varphi$ is consistent.*

Proof. Suppose for a contradiction that $T \subseteq \mathcal{L}$ and satisfiable $\varphi \in \mathcal{L}$ are such that $T \oplus_{\mathcal{D}} \varphi$ is inconsistent. Since $\mathcal{D}$ is valid, we infer by compactness that there exists finite $D \subseteq T \cup \{\varphi\}$ such that $D \subseteq T \oplus_{\mathcal{D}} \varphi$, $D \setminus \{\varphi\} \neq \emptyset$, and D is minimally inconsistent. Then there exists a $\mathcal{D}$-proof of $\bot$ from D such that every member of D labels some of its leaves, which implies that some member of $D \setminus \{\varphi\}$ occurs in an e.c. $\mathcal{D}$-proof from $T \cup \{\varphi\}$. This contradicts the fact that $D \subseteq T \oplus_{\mathcal{D}} \varphi$. $\square$

When we do introduce a relevance relation that is non-trivial, we would like to obtain the same property. This is indeed the case when the initial theory along with the new information is a) neatly partitioned by the relevance relation and b) does not have any of its implicit logical consequences made explicit by the revision process.

Proposition 3 *Let $T, T' \subseteq \mathcal{L}$ and $\varphi \in \mathcal{L}$ be such that $T' = \{\psi \in T \mid \varphi$ does not τ-trigger $\psi\}$. Suppose that the following holds.*

- *$\mathcal{D}$ is a derivation frame all of whose members are trees of the form*



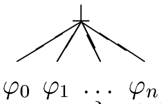
  with $\{\varphi_0 \ldots \varphi_n\}$ inconsistent;
- *for all minimally inconsistent subsets D of $\mathcal{L}$, there exists a member of $\mathcal{D}$ whose set of leaves contains D;*
- *$T \cup \{\varphi\}$ is τ-partitioned.*

If T' is consistent and φ is satisfiable, then $T \oplus^{\tau}_{\mathcal{D}} \varphi$ is consistent.

Proof. Trivially, $T \oplus^{\tau}_{\mathcal{D}} \varphi = T' \cup (T \setminus T') \oplus^{\tau}_{\mathcal{D}} \varphi$. By the definition of $\mathcal{D}$, $(T \setminus T') \oplus^{\tau}_{\mathcal{D}} \varphi \subseteq (T \setminus T') \cup \{\varphi\}$. Since $T \cup \{\varphi\}$ is τ-partitioned, no formula in T' is τ-triggered by φ, and every member of $T \setminus T'$ is τ-triggered by φ, it is easy to verify that for all $\psi \in T'$ and $\chi \in (T \setminus T') \cup \{\varphi\}$, neither ψ τ-triggers χ nor χ τ-triggers ψ. Hence by Proposition 1, it suffices to show that $(T \setminus T') \oplus^{\tau}_{\mathcal{D}} \{\varphi\}$ is consistent. This follows immediately from Proposition 2. $\square$

When we do introduce a non-trivial relevance relation and more powerful derivation frames, the issue is more complicated and is deferred for the time being to a more leisurely investigation. The following shows that full meet *base* revision [8, Definition 1.4] is easily and naturally obtained in our framework.

Proposition 4 *Suppose that $\mathcal{D}$ is the derivation frame consisting of all proof trees of form*

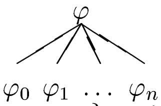


with $\{\varphi_0 \dots \varphi_n\}$ minimally inconsistent. Then for all $T \subseteq \mathcal{L}$ and satisfiable $\varphi \in \mathcal{L}$, $T \oplus_{\mathcal{D}} \varphi$ is equal to $T \dotplus \varphi$, where $\dotplus$ is the full meet revision operator for base revision.

Proof. Let $T \subseteq \mathcal{L}$ and satisfiable $\varphi \in \mathcal{L}$ be given. By [8, Observation 2.10], the full meet contraction of T by $\neg\varphi$ for base revision is the set of all $\psi \in T$ which do not belong to a minimally inconsistent subset of $T \cup \{\varphi\}$. The result follows immediately. $\square$

Given the result above, we might expect a similar connection between our framework and full meet AGM revision (*i.e.*, if $\neg\varphi \in T$, then $T * \varphi$ is the logical closure of $\{\varphi\}$ where $*$ is the full meet revision operator). Such a connection exists, but as the following shows, it is not perfect. The derivation frame considered below is the counterpart of the one in Proposition 4: it permits not just the detection of inconsistencies but the derivation of arbitrary consequences of a given set of formulas. For this derivation frame $\mathcal{D}$, $T \oplus_{\mathcal{D}} \varphi$ is logically equivalent but not equal to $T * \varphi$. The formula φ occurs in both sets, but some of its logical consequences which necessarily appear in $T * \varphi$ can remain merely implicit in $T \oplus_{\mathcal{D}} \varphi$. This is because AGM revision forces closure at the final stage of the revision process.

Proposition 5 *Suppose that $\mathcal{D}$ is the derivation frame generated by the set of all proof trees of form*



with $\{\varphi_0 \dots \varphi_n\}$ a minimal subset of $\mathcal{L}$ that implies φ (including the case where $\varphi = \perp$ and $\{\varphi_0 \dots \varphi_n\}$ is minimally inconsistent). For all $T \subseteq \mathcal{L}$ and satisfiable $\varphi \in \mathcal{L}$, if $T \cup \{\varphi\}$ is inconsistent then $T \oplus_{\mathcal{D}} \varphi$ is a subset of the set of all logical consequences of φ.

Proof. Let $T \subseteq \mathcal{L}$ and satisfiable $\varphi \in \mathcal{L}$ be such that $T \cup \{\varphi\}$ is inconsistent. By the definition of $\mathcal{D}$, the set Z of all logical consequences of $T \cup \{\varphi\}$ is equal to the set of all members of $\mathcal{L}$ that occur in at least one $\mathcal{D}$-proof from $T \cup \{\varphi\}$. Let $\psi \in Z$ be such that $\varphi \not\models \psi$. It suffices to show that $\psi \notin T \oplus_{\mathcal{D}} \varphi$. Using an argument similar to the one used in the proof of [8, Observation 2.12], it is easy to verify that ψ belongs to a minimally inconsistent subset D of Z. We can choose a minimal subset E of $T \cup \{\varphi\}$ such that $E \models D$. We infer immediately from the definition of a derivation frame that all members of D occur in some e.c. $\mathcal{D}$-proof from E, hence that ψ occurs in some e.c. $\mathcal{D}$-proof from $T \cup \{\varphi\}$. Hence $\psi \notin T \oplus_{\mathcal{D}} \varphi$, as wanted. $\square$

4 Related Work and Conclusion

A detailed investigation of syntactic and semantic notions of relevance is presented in [9]. A development of a probability inspired 3-place notion of relevance among sets of propositional atoms occurs in [5]. A comprehensive collection of papers dealing with relevance may be found in [1]. A permissive model of belief revision that retains weakenings of formulas is presented in [4] while an implementation of a AGM-style revision system using first-order theorem provers may be found in [6]. Our model incorporates the insights offered in [11] and [3]. These models offered a syntactic definition of relevance based on sharing of non-logical symbols between formulas and used it to guide the revision process. An extension to a first-order setting was left as an open problem.

We have presented a general model for first-order belief revision characterized by an underlying relevance relation and a background proof system. A wide variety in these characterizing parameters is easily accomodated. It allows weakenings of beliefs which were initially implicit to become explicit and survive the revision process. The usage of deductively limited proof systems permits an inconsistency tolerant model. In future work, we will investigate the properties of proof systems $\mathcal{D}$ such that given a finite set of sentences T, the set of all logical consequences that can be generated from T on the basis of $\mathcal{D}$ is also finite. As well, we would like to limit explicit contradictions to those derived from a pair of formulas of the form $\varphi, \neg\varphi$. Such a proof system will enjoy the facility of being easily implemented and would allow the resolution of some forms of inconsistency. Moreover, it should be specifically first-order in that it should contain rules for quantified formulas. We believe ourselves to have provided the foundations for such a task.

References

1. AIJ. Special issue on relevance. In *Artificial Intelligence*, volume 97(1-2). 1997.
2. Carlos Alchourron, Peter Gärdernfors, and David Makinson. On the logic of theory change: Partial meet functions for contraction and revision. *J. of Symbolic Logic*, 50:510–530, 1985.
3. Samir Chopra and Rohit Parikh. An inconsistency tolerant model for belief representation and belief revision. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 192–197. Morgan-Kaufmann, 1999.
4. Maria Cravo, Joao Cachopo, Ana Cachopo, and Joao Martins. Permissive belief revision (preliminary report). In *Proceedings of NRAC-99*, 1999.
5. Adnan Darwiche. A logical notion of conditional independence: properties and applications. *Artificial Intelligence*, 97(1-2)(8):45–82, 1997.
6. S. E. Dixon and W. R. Wobcke. The implementation of a first-order logic AGM belief revision system. In *Proceedings of 5th IEEE International Conference on Tools with AI*, pages 40–47. IEEE Computer Society Press, 1993.
7. Sven Ove Hansson. A dyadic representation of belief. In Peter Gärdenfors, editor, *Belief Revision*, Theoretical Computer Science. Cambridge University Press, Cambridge, 1992.
8. Sven Ove Hansson. *A textbook of belief dynamics*. Kluwer Academic Publishers, 1999.
9. Jerome Lang and Pierre Marquis. Complexity results for independence and definability in propositional logics. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference*, pages 356–367, San Mateo, CA, 1998. Morgan-Kaufmann.
10. Bernhard Nebel. Syntax based approaches to belief revision. In Peter Gärdenfors, editor, *Belief Revision*, Theoretical Computer Science. Cambridge University Press, 1992.
11. Rohit Parikh. Beliefs, belief revision, and splitting languages. In Lawrence Moss, Jonathan Ginzburg, and Maarten de Rijke, editors, *Logic, Language, and Computation, Volume 2*, CSLI Lecture Notes No. 96, pages 266–268. CSLI Publications, 1999.