Specifying Plausibility Levels for Iterated Belief Change in the Situation Calculus

Toryn Q. Klassen and **Sheila A. McIlraith** and **Hector J. Levesque**
Department of Computer Science, University of Toronto, Canada
{toryn, sheila, hector}@cs.toronto.edu

Abstract

We investigate augmenting a theory of belief and actions with qualitative plausibility levels. Shapiro et al. created a framework for modeling iterated belief revision and update which integrated those features with the well-developed theory of action in the situation calculus. However, applying their technique requires associating plausibility levels with initial situations, for which no very convenient mechanism had been proposed. Schwering and Lakemeyer proposed deriving these initial plausibility levels from a set of conditionals, similarly to how models are ranked in Pearl's System Z. However, their approach inherits some limitations of System Z. We consider alternatives, and argue that a perspicuous approach is to measure plausibility by counting the abnormalities in a situation (similarly to cardinality-based circumscription). By allowing abnormalities to change over time, we can also model changing plausibility levels in a natural and simple way, which gives us a flexible approach for handling belief change about predicted and unpredicted exogenous actions.

1 Introduction

How an agent's beliefs should change over time is a much-studied area in artificial intelligence (for example, see Peppas (2008) for an overview of work in belief revision). In this paper, we present a framework supporting (1) iterated belief revision and update and (2) the modeling of action and change, in the context of (3) a simple qualitative specification of what the agent considers plausible. To do so, we build on the work of Shapiro et al. (2011), who created a framework for modeling iterated belief change in the situation calculus (McCarthy and Hayes 1969; Reiter 2001). Their approach already has properties (1) and (2); to achieve (3), we incorporate a way of specifying levels of plausibility.

Let us first explain the relevance of plausibility to the framework. A central idea behind Shapiro et al.'s approach to belief change is that the agent's beliefs are determined by truth in all the *most plausible* accessible situations, and it is the accessibility relation, not the plausibility levels, that changes over time. (This idea is also used in the modal logic of Friedman and Halpern (1999).) Shapiro et al.'s approach is integrated into the well-developed theory of action that exists for the situation calculus, including Reiter's solution to the frame problem (Reiter 2001).

However, the initial plausibility levels still have to be described somehow, which has been viewed as difficult. To avoid using plausibility levels at all, Demolombe and Parra (2006) even created an alternative approach to belief revision that instead had sensing actions modify "imaginary" situations that were accessible to agents. Schwering and Lakemeyer (2014), on the other hand, proposed to build on Shapiro et al.'s framework by adding a mechanism ("only-believing") for automatically deriving plausibility levels from a set of conditionals. This derivation is essentially the same as the one used by System Z (Pearl 1990), a system for default reasoning, in ranking models based on conditionals. As we will show later, Schwering and Lakemeyer's approach inherits some limitations of System Z. In particular, the derivation procedure does not treat the conditionals as being "independent" of each other, which makes specifying some natural plausibility orderings more difficult than might be expected. Furthermore, only a finite number of distinct plausibility levels can be defined using this approach (at most, roughly the same number of levels as the number of conditionals used in defining them).

In this paper, we apply other ideas from the nonmonotonic reasoning literature to the problem of specifying plausibility levels for use in Shapiro et al.'s framework. One approach would be to replace Schwering and Lakemeyer's use of System Z with another mechanism from conditional logics in the literature, such as lexicographic entailment (Benferhat et al. 1993; Lehmann 1995). We do make some remarks on this later, but we instead suggest an approach based on *cardinality-based circumscription*. We show that it is in a sense more general than lexicographic entailment.

Circumscription (McCarthy 1980; 1986; Lifschitz 1994) is one of the most widely-studied approaches to nonmonotonic reasoning. It involves minimizing the extensions of certain predicates (e.g., *abnormality* predicates). Minimality is traditionally defined in terms of set inclusion, but in *cardinality-based* circumscription (CBC) (Liberatore and Schaerf 1995; 1997; Sharma and Colomb 1997; Moinard 2000), the extensions of predicates are compared by cardinality instead. In the context of belief revision, Klassen, Levesque, and McIlraith (2017), in their modal logic based on Friedman and Halpern's, used the cardinal-

ities of abnormality predicates to determine the plausibility levels of worlds (though they did not give much argument for this design choice). We suggest applying the same idea in the framework of Shapiro et al., and demonstrate with some examples the efficacy of this approach.

Furthermore, while in the original framework of Shapiro et al. plausibility levels were fixed, it is natural to consider allowing ordinary non-sensing actions to change the extensions of abnormality predicates. This turns out to provide a simple way of representing the plausibility of exogenous events, which is more general than a previous extension of the framework of Shapiro et al. to exogenous events that was proposed by Shapiro and Pagnucco (2004).

The outline of this paper follows. First, in §2 we review the relevant background. Then in §3, we show how to use the cardinalities of abnormality predicates along with *only-knowing* (Lakemeyer and Levesque 1998) to specify plausibility levels in the situation calculus, and show the advantages of this over Schwering and Lakemeyer's approach. In §4, we relax the condition that abnormalities have to be unchanging, and develop our new approach to modeling the plausibility of situations whose histories include exogenous actions. In §5, we justify our choice of CBC by comparing it with lexicographic entailment. We also explain why regular circumscription would not have been a suitable choice. Finally, after a comparison with related work (§6), we conclude in §7 with suggestions for future work.

2 Preliminaries

We start by briefly reviewing the situation calculus, and then the approach to iterated belief change in the situation calculus due to Shapiro et al. (2011). We then consider how Schwering and Lakemeyer (2014) extended that approach with only-believing, and limitations of that extension.

2.1 The situation calculus

The situation calculus (McCarthy and Hayes 1969; Reiter 2001) is a predicate calculus language for describing action and change. It is a typed language, with sorts for *situations*, *actions*, and *objects* (everything else – for convenience, we will have the natural numbers as a subsort of objects).

We now describe some notational conventions. Each of the sorts of symbols we describe below may also appear with decorations (e.g., subscripts). We will use s and t as variables of type situation; a as a variable of type action; i and j as numeric variables; and x, y, and z as variables for objects. Predicate symbols start with an uppercase letter, and function/constant symbols with a lowercase letter (except for the constant symbol $\mathtt{S}_0$, denoting the actual initial situation, and for some use of infix notation). We will use uppercase Roman letters like P and Q for predicate variables, lowercase Greek letters like ϕ and ψ for formulas, and uppercase Greek letters like Σ for sets of formulas. We will sometimes leave outer universal quantifiers on sentences implicit, e.g., using $\phi(x)$ to stand for $(\forall x).\phi(x)$, though not for higher-order quantifiers. We may abbreviate sequences of quantified variables with vectors, e.g., using $(\forall \vec{x})$ for $(\forall x_1, x_2, \ldots, x_k)$.

In the situation calculus, situations represent histories of actions performed starting from an initial situation. The initial situations represent different possible ways the world could start out; as noted previously, we will use the term $\mathtt{S}_0$ to denote the actual initial situation. The situation that results from performing action a in situation s is written $\mathtt{do}(a, s)$. We may abbreviate the application of successive actions $a_1, \ldots, a_k$ in situation s by writing $\mathtt{do}([a_1, \ldots, a_k], s)$. We will use the abbreviation $\mathtt{Init}(s) \stackrel{\text{def}}{=} \neg(\exists a, s').s = \mathtt{do}(a, s')$ to say that s is an initial situation. The binary predicate $s \sqsubset s'$ indicates that situation s' is the result of performing one or more actions in s. Properties that actions can change are described using *fluents*, which are predicates (or functions) whose last argument is a situation term.

The standard way of axiomatizing domains in the situation calculus is by using some variation of *basic action theories* (Reiter 2001). A basic action theory consists of the following sets of axioms: initial state axioms, which describe the initial situations; precondition axioms that describe when actions are possible to execute[1]; successor state axioms (SSAs), specifying for each fluent how its value in a non-initial situation depends on the previous situation; sensing axioms (more on these below); unique names axioms for actions; and domain-independent foundational axioms (including a second-order induction axiom).

An agent's beliefs can be described in the situation calculus, following Scherl and Levesque (2003), by modeling the epistemic accessibility relation with a fluent $\mathtt{B}(s', s)$ that says that s' is epistemically accessible from s (note the order of the arguments). A special predicate $\mathtt{SF}(a, s)$, described by the sensing axioms, determines the (binary) sensing result the agent gets from performing action a in s. The accessibility relation $\mathtt{B}$ has the following SSA:

$$\mathtt{B}(s'', \mathtt{do}(a, s)) \equiv (\exists s').\mathtt{B}(s', s) \wedge s'' = \mathtt{do}(a, s') \wedge (\mathtt{SF}(a, s') \equiv \mathtt{SF}(a, s)).$$

That is, for a situation to be accessible after performing an action a, that situation must the result of doing a in some other situation that was previously accessible, and the sensing result of a must reflect the true value. In Scherl and Levesque's approach, what was believed was defined by what was true in all the accessible situations.

2.2 Iterated belief change in the situation calculus

Shapiro et al. (2011) created a framework for iterated belief change with the situation calculus. They used the $\mathtt{B}$ and $\mathtt{SF}$ predicates described above, but in order to allow for beliefs to be retracted (which Scherl and Levesque did not), Shapiro et al. defined belief as truth in the *most plausible* accessible situations rather than in all accessible situations. With this approach, sensing can cause an agent to lose a belief by making inaccessible all the situations that were previously the most plausible accessible ones.

Shapiro et al. used a function $\mathtt{pl}$ to assign plausibility levels (natural numbers) to situations. The SSA for $\mathtt{pl}$ specifies that the function never changes: $\mathtt{pl}(\mathtt{do}(a, s)) = \mathtt{pl}(s)$.

[1] For this paper we will assume, like Shapiro et al. (2011), that all actions are always executable, so will not refer to these again.

Belief was defined in terms of plausibility and accessibility. First, Shapiro et al. defined $\texttt{MP}(s', s)$ to mean that s' is at least as plausible as any situation accessible from s:

Definition 1 (Shapiro et al. (2011, Definition 10))**.**

$$\texttt{MP}(s', s) \stackrel{\text{def}}{=} (\forall s'').\texttt{B}(s'', s) \supset \texttt{pl}(s') \leq \texttt{pl}(s'')$$

Then they introduced $\texttt{MPB}(s', s) \stackrel{\text{def}}{=} \texttt{B}(s', s) \wedge \texttt{MP}(s', s)$, to mean that s' is one of the most plausible situations accessible from s. Finally, they used MPB in defining a belief operator Bel:

$$\texttt{Bel}(\phi, s) \stackrel{\text{def}}{=} (\forall s').\texttt{MPB}(s', s) \supset \phi[s']$$

Here ϕ is a formula which may contain the special situation term now (meant to indicate the current situation), and $\phi[s']$ is the result of substituting s' for now in ϕ. So $\texttt{Bel}(\phi, s)$ is true if $\phi[s']$ is true in the most plausible accessible situations from s. For readability, we may (like Shapiro et al.) suppress the now argument of fluents within the scope of a belief operator, e.g., writing $\texttt{Bel}(F(\vec{x}), s)$ for $\texttt{Bel}(F(\vec{x}, \texttt{now}), s)$.

They used the SSA for B that we have noted in §2.1. There also are some initial state axioms relating to B (for introspection, and making the agent initially believe it's in an initial situation): $\texttt{Init}(s) \wedge \texttt{B}(s', s) \supset (\forall s'').\texttt{B}(s'', s') \equiv \texttt{B}(s'', s)$ and $\texttt{Init}(s) \wedge \texttt{B}(s', s) \supset \texttt{Init}(s')$.

They showed that their approach satisfies a slightly modified version of the well-known AGM postulates for belief revision (Alchourrón, Gärdenfors, and Makinson 1985). It also satisfies some of the KM postulates for belief update (Katsuno and Mendelzon 1991), and some of the DP postulates for iterated belief revision (Darwiche and Pearl 1997).

2.3 Deriving plausibilities using only-believing

Writing initial state axioms to explicitly assign plausibility levels can be inconvenient. As Schwering and Lakemeyer (2014) (and even Shapiro et al. themselves) point out, the actual numbers used for plausibility levels are not very important. We may also note that writing explicit numbers in an action theory may make it harder to modify.

A proposal to address this problem can be found in the logic $\mathcal{ESB}$ (Schwering and Lakemeyer 2014; Schwering, Lakemeyer, and Pagnucco 2017). In $\mathcal{ESB}$, which is a modal version of situation calculus without any explicit situation terms, initial plausibility levels can be determined from a set of conditionals that are "only-believed".[2]

In the semantics of $\mathcal{ESB}$, an *epistemic state* is a sequence $e = (e_1, e_2, e_3, \dots)$ of sets of *worlds*, where $e_j \subseteq e_{j+1}$ (and the sequence converges, in that for some N, $e_n = e_N$ when $n > N$). For the purposes of this paper, it will suffice to understand a world as providing truth values for first-order sentences (we will not go over how actions are handled in $\mathcal{ESB}$). The idea is that the entries $e_1, e_2, e_3, \dots$ in an epistemic state e correspond to plausibility levels, with less plausible worlds only being in higher-numbered entries. An epistemic state e satisfies the sentence $\mathbf{B}(\phi \Rightarrow \psi)$ if ψ is true at all the most plausible worlds where ϕ is true. Note that the conditional '$\Rightarrow$' is distinct from the material conditional, '$\supset$'. Belief in ϕ can be defined with $\mathbf{B}(\texttt{True} \Rightarrow \phi)$.

Now, let us explain *only-believing*. Suppose that $\Gamma = \{\phi_1 \Rightarrow \psi_1, \dots, \phi_m \Rightarrow \psi_m\}$, where each ϕ_i and ψ_i is an objective formula (not containing any belief or knowledge operators). Let Γ' be the set $\{\phi_1 \supset \psi_1, \dots, \phi_m \supset \psi_m\}$ that is like Γ but with the conditional symbols replaced by material conditionals. The semantics of only-believing is as follows: an epistemic state $e = (e_1, e_2, e_3, \dots)$ satisfies the sentence $\mathbf{O}(\Gamma)$ ("Γ is all that is believed") iff e_1 satisfies all the material conditionals in Γ' and e_{j+1} satisfies the subset of those material conditionals whose antecedents are not true in any world in e_j. That is, $e_1 = \{w : w \models \bigwedge \Gamma'\}$ and for each $j \geq 1$ we have that e_{j+1} is equal to $\{w : w \models \bigwedge\{(\phi_i \supset \psi_i) \in \Gamma' : \forall w' \in e_j, w' \not\models \phi_i\}\}$. This ordering on worlds is essentially that given by System Z (Pearl 1990), as described by Schwering (2016, §4.7).

2.4 Issues with only-believing

Schwering and Lakemeyer's only-believing is similar to the propositional System Z, and inherits some limitations as a result of this, as Schwering, Lakemeyer, and Pagnucco (2017, p. 75) note.

In particular, the conditionals that are only-believed are not treated as being fully "independent" of each other. Adapting an example from Pearl (1990, §3), we have

$$\mathbf{O}(\text{PENGUIN} \Rightarrow \text{BIRD}, \text{BIRD} \Rightarrow \text{FLY}, \text{PENGUIN} \Rightarrow \neg\text{FLY}, \text{BIRD} \Rightarrow \text{BEAK}) \models \neg\mathbf{B}(\text{PENGUIN} \Rightarrow \text{BEAK})$$

An intuitive reading of what's believed is that a penguin most plausibly is a bird (PENGUIN $\Rightarrow$ BIRD), a bird most plausibly flies (BIRD $\Rightarrow$ FLY), a penguin most plausibly doesn't fly (PENGUIN $\Rightarrow \neg$FLY), and a bird most plausibly has a beak (BIRD $\Rightarrow$ BEAK). With these beliefs, the agent unfortunately does not believe that a penguin most plausibly has a beak (PENGUIN $\Rightarrow$ BEAK). This has been called the "drowning problem", and what is lacking from System Z and other systems with this problem has been called "strong independence" (Strasser and Antonelli 2016).

To give perhaps the simplest example that shows the problem, the epistemic state corresponding to $\mathbf{O}(\texttt{True} \Rightarrow P, \texttt{True} \Rightarrow Q)$ – that is, to only-believing that P is most plausibly true and that Q is most plausibly true – has only two distinct entries, $e_1 = \{w : w \models P \wedge Q\}$ and $e_2 = e_3 = \dots$ is the set of all worlds. If the agent with this epistemic state were to learn that P were false, on revising their beliefs they would also lose their belief in Q (since they would discard all worlds from e_1). Intuitively, we would like to have that P and Q are features that independently contribute to the plausibility of a world.

Aside from the lack of strong independence, another issue with only-believing is that despite being used in a first-order logic, it works essentially the same as the propositional System Z. The epistemic state induced by only-believing a finite number m of conditionals will only have a finite number of distinct entries – at most $m + 1$ (Schwering 2016, Theorem 4.5.3). However, it's easy to come up with examples for which it's desirable to distinguish between a number of plausibility levels that does not have a clear bound. For example,

[2] Shapiro et al. also considered using conditionals to constrain plausibility levels, though not in a way that defined them uniquely.

for every n, an agent might think that a conspiracy involving n people is more plausible than one with $n+1$ people (we will formalize this example in §3.4).

3 The approach

In this section we develop our alternative for specifying plausibility levels. Conditional logics are not the only way to do non-monotonic reasoning, nor to assign plausibility levels. Another popular form of non-monotonic reasoning is based on circumscription (McCarthy 1980; 1986; Lifschitz 1994). Here we will consider a variant of circumscription, *cardinality-based* circumscription (CBC) (Liberatore and Schaerf 1995; 1997; Sharma and Colomb 1997; Moinard 2000). After describing CBC and showing how it can be expressed in second-order logic, we will show how we can use it as the basis for determining plausibility levels in the situation calculus by slightly modifying Shapiro et al.'s action theories. As our examples will demonstrate, unlike only-believing, CBC avoids the drowning problem and can specify an infinite number of distinct plausibility levels.

3.1 Cardinality-based circumscription

Here we present a simple form of prioritized CBC, based on the work of Klassen, Levesque, and McIlraith (2017), where prioritized *abnormality* predicates (McCarthy 1986) are minimized and no predicates are kept fixed.[3] We will be using the abnormality predicates as a way of measuring plausibility (which may lead one to want to write slightly different theories than if they were instead measuring *typicality*, though we will not discuss this distinction further).

Suppose that we have a finite set of abnormality predicates $\mathtt{Ab}_1, \mathtt{Ab}_2, \ldots$, each with an associated priority (intuitively, a higher priority abnormality is a sign of greater implausibility). Let us say that there are k distinct priority levels, and that $\vec{A}^i$ is the list of abnormality predicates of the ith highest priority. To any interpretation $\mathfrak{I} = \langle \mathcal{D}, \mathcal{I} \rangle$, with domain $\mathcal{D}$ and interpretation mapping $\mathcal{I}$, we can assign a k-ary *abnormality vector* $\vec{c}(\mathfrak{I})$ where each entry is either a natural number or ∞, and whose ith entry is the sum of the cardinalities of the extensions of the priority i abnormality predicates, i.e. $\vec{c}(\mathfrak{I})_i = \sum_{\mathtt{Ab} \in \vec{A}^i} |\mathcal{I}[\mathtt{Ab}]|$. Note that we do not distinguish between different infinite cardinalities (i.e., there is only one ∞), and that for a 0-ary predicate, the cardinality of its extension will either be 0 or 1.

Abnormality vectors can be ordered in a lexicographic way, i.e., we define $\vec{c}(\mathfrak{I}_1) < \vec{c}(\mathfrak{I}_2)$ if there is some i so that $c(\mathfrak{I}_1)_i < c(\mathfrak{I}_2)_i$ and so that for all $j < i$, we have $c(\mathfrak{I}_1)_j \leq c(\mathfrak{I}_2)_j$. That is, lesser abnormality vectors are ones that count a smaller number of abnormalities, giving higher priority to the higher priority abnormalities. We can then define (as usual for circumscription) a form of entailment in which only the minimal models are considered (note that since the abnormality vectors are well-ordered, there are never infinite descending chains of models).

[3] Being able to keep some predicates fixed has uses, e.g., to prevent minimizing the number of abnormally non-flying birds from minimizing the number of penguins, but it also introduces complications (Brachman and Levesque 2004, §11.3.3).

Definition 2. For Δ a set of sentences and β a sentence, we write $\Delta \models_{\mathsf{card}} \beta$ if for every interpretation $\mathfrak{I}$ such that $\mathfrak{I} \models \Delta$, either $\mathfrak{I} \models \beta$ or there is another interpretation $\mathfrak{I}'$ such that $\vec{c}(\mathfrak{I}') < \vec{c}(\mathfrak{I})$ and $\mathfrak{I}' \models \Delta$.

To give a classic example (which is simple enough that CBC behaves like regular circumscription on it), we have that $\{\text{BIRD} \wedge \neg\mathtt{Ab} \supset \text{FLY}, \text{BIRD}\} \models_{\mathsf{card}} \text{FLY}$. That is, if there is a bird, and the bird flies unless $\mathtt{Ab}$ is true, then the bird is assumed to fly. This is because in the minimal models, the cardinality of the extension of $\mathtt{Ab}$ is minimized.

3.2 Expressing CBC in second-order logic

As for regular circumscription, it's also possible to describe CBC using formulas of second-order logic. This was shown for some forms of CBC by Sharma and Colomb (1997, §4.1.1), and we can do the same for ours, based on their approach. Some of this machinery will be useful when we turn to incorporating counting abnormalities into the situation calculus.

Suppose that $\vec{P} = \langle P_1, \ldots, P_m \rangle$ and $\vec{Q} = \langle Q_1, \ldots, Q_m \rangle$ are lists of predicates. The sum of cardinalities of the extensions of $P_1, \ldots, P_m$ is at most the sum of the cardinalities of the extensions of $Q_1, \ldots, Q_m$ iff there is an injective function from the disjoint union of the extensions of $P_1, \ldots, P_m$ to the disjoint union of the extensions of $Q_1, \ldots, Q_m$. We can express this property in second order logic (Sharma and Colomb did so for the case where $m = 1$).

Definition 3. We will use the abbreviation $\vec{P} \leq_{\mathsf{card}} \vec{Q}$ to stand for the second-order sentence

$$(\exists F_{ij} : 1 \leq i, j \leq k).\mathtt{INJECTIVE}(F_{ij} : 1 \leq i, j \leq k) \wedge \bigwedge_i \Big[(\forall \vec{x}_i).P_i(\vec{x}_i) \supset \bigvee_j (\exists \vec{y}_j)\Big(F_{ij}(\vec{x}_i, \vec{y}_j) \wedge Q_j(\vec{y}_j)\Big)\Big]$$

where $\mathtt{INJECTIVE}(F_{ij} : 1 \leq i, j \leq k)$ is an abbreviation for

$$\bigwedge_{i,j} (\forall \vec{x}_i, \vec{x}'_i, \vec{y}_j)[F_{ij}(\vec{x}_i, \vec{y}_j) \wedge F_{ij}(\vec{x}'_i, \vec{y}_j) \supset \vec{x}_i = \vec{x}'_i] \wedge \bigwedge_{i,j,k:i \neq k} (\forall \vec{x}_i, \vec{x}_k, \vec{y}_j) \neg [F_{ij}(\vec{x}_i, \vec{y}_j) \wedge F_{kj}(\vec{x}_k, \vec{y}_j)].$$

Now, this $\leq_{\mathsf{card}}$ relation compares predicates by cardinality, but in a way that is a bit more fine-grained than what we want, since it discriminates between differing infinite cardinalities – unlike the abnormality vectors we defined earlier. To match those, we want to define a relation that is like $\leq_{\mathsf{card}}$ except for treating all infinities as being equal.

To do so, let us first define that $\mathtt{INF}(P)$, where P is a predicate symbol, abbreviates the second-order sentence

$$(\exists R).(\forall \vec{x}, \vec{y}, \vec{z})[R(\vec{x}, \vec{y}) \wedge R(\vec{y}, \vec{z}) \supset R(\vec{x}, \vec{z})] \wedge (\forall \vec{x})[\neg R(\vec{x}, \vec{x}) \wedge (\exists \vec{y}) P(\vec{y}) \wedge R(\vec{x}, \vec{y})],$$

saying that there is a transitive, irreflexive, serial relation on the extension of P. This is true iff P has an infinite extension. Note that the number of entries in each of $\vec{x}, \vec{y}$, and $\vec{z}$ in the expansion of $\mathtt{INF}(P)$ matches the arity of P. Finally, we can define a relation $\leq^{\infty}_{\mathsf{card}}$ that is like $\leq_{\mathsf{card}}$ except for treating all infinities as being equal.

Definition 4. We define $\vec{P} \leq^{\infty}_{\mathsf{card}} \vec{Q}$ as the sentence

$$(\vec{P} \leq_{\mathsf{card}} \vec{Q}) \vee \bigvee_{i,j} \big(\mathtt{INF}(P_i) \wedge \mathtt{INF}(Q_j)\big).$$

We also define $\vec{P} <^{\infty}_{\mathsf{card}} \vec{Q}$ as $\neg(\vec{Q} \leq^{\infty}_{\mathsf{card}} \vec{P})$.

Finally, we want to define a relation that treats some predicates as higher priority than others. Suppose that we partition the elements of $\vec{P}$ among $\vec{P}^1, \ldots, \vec{P}^k$ (where $k \leq m$), so that $\vec{P}^1$ contains the highest priority predicates from $\vec{P}$, $\vec{P}^2$ contains the second highest priority predicates, and so on. Then we define the prioritized relation $\prec^{\infty}_{\mathsf{card}}$ as follows:

Definition 5. Let $\vec{P}^1, \ldots, \vec{P}^k \prec^{\infty}_{\mathsf{card}} \vec{Q}^1, \ldots, \vec{Q}^k$ abbreviate

$$\bigvee_i \Big(\vec{P}^i <^{\infty}_{\mathsf{card}} \vec{Q}^i \wedge \Big(\bigwedge_{j<i} \vec{P}^j \leq^{\infty}_{\mathsf{card}} \vec{Q}^j \Big) \Big).$$

We can then also define $\vec{P}^1, \ldots, \vec{P}^k \preccurlyeq^{\infty}_{\mathsf{card}} \vec{Q}^1, \ldots, \vec{Q}^k$ as $\neg(\vec{Q}^1, \ldots, \vec{Q}^k \prec^{\infty}_{\mathsf{card}} \vec{P}^1, \ldots, \vec{P}^k)$.

Finally, given a sentence α, it is possible to use $\prec^{\infty}_{\mathsf{card}}$ to define a second-order sentence that entails β just in case $\alpha \models_{\mathsf{card}} \beta$. For completeness, the details of this are given below, but they will not be needed in subsequent sections.

Definition 6. Let $\vec{P}^1, \ldots, \vec{P}^k$ be disjoint lists of predicate symbols and $\vec{Z}$ a list of other predicate (and/or function) symbols. The prioritized circumscription of $\vec{P}^1, \ldots, \vec{P}^k$ in α with varied $\vec{Z}$ is the second-order sentence

$$\alpha(\vec{P}^1, \ldots, \vec{P}^k, \vec{Z}) \wedge \neg(\exists \vec{Q}^1, \ldots, \vec{Q}^k, \vec{Z}').[\, \alpha(\vec{Q}^1, \ldots, \vec{Q}^k, \vec{Z}') \wedge \vec{Q}^1, \ldots, \vec{Q}^k \prec^{\infty}_{\mathsf{card}} \vec{P}^1, \ldots, \vec{P}^k],$$

which we will denote by $\mathrm{NCIRC}[\alpha; \vec{P}^1 > \cdots > \vec{P}^k; \vec{Z}]$.

Note that the models of $\mathrm{NCIRC}[\alpha; \vec{P}^1 > \cdots > \vec{P}^k; \vec{Z}]$ are those that make $\alpha(\vec{P}^1, \ldots, \vec{P}^k, \vec{Z})$ true while minimizing (in a prioritized, cardinality-based way) the extensions of the predicates in $\vec{P}^1, \ldots, \vec{P}^k$ (even if the minimization requires changing the interpretations of the elements of $\vec{Z}$).

Proposition 1. *Let $\vec{Z}$ include all predicate/function symbols, other than the abnormality predicates in $\vec{A}^1, \ldots, \vec{A}^k$. Then $\alpha \models_{\mathsf{card}} \beta$ iff $\mathrm{NCIRC}(\alpha; \vec{A}^1 > \cdots > \vec{A}^k; \vec{Z}) \models \beta$.*

3.3 Determining the plausibility of situations

We now return to discussing the situation calculus. In order to compare the plausibility levels of situations, we propose to introduce *abnormality fluents*. Each abnormality fluent keeps the same value over time, as specified by SSAs of the form $\mathtt{Ab}_i(\vec{x}, \mathtt{do}(a, s)) \equiv \mathtt{Ab}_i(\vec{x}, s)$ for each i. Later on (in §4) we will relax this condition, but for now we are following the approach of Shapiro et al., where plausibility levels do not change.

There are priorities associated with the abnormality fluents. Let us use the notation $\vec{A}^i[s]$ to refer to the list of priority i abnormality fluents, with their situation terms fixed to s. Rather than explicitly assigning plausibility values based on abnormalities, it's simpler to just redefine $\mathtt{MP}(s', s)$, which was used in the definition of `Bel` in determining whether s' is as plausible as any situation accessible from s.

Definition 7 (redefining `MP` (from Definition 1))**.**

$$\mathtt{MP}(s', s) \stackrel{\text{def}}{=} (\forall s'').\big(\mathtt{B}(s'', s) \supset \vec{A}^1[s'], \ldots, \vec{A}^k[s'] \preccurlyeq^{\infty}_{\mathsf{card}} \vec{A}^1[s''], \ldots, \vec{A}^k[s'']\big)$$

Where before the plausibility of s' and s'' was compared by comparing $\mathtt{pl}(s')$ and $\mathtt{pl}(s'')$, now we check in which situation more abnormal fluents hold. All the rest of the machinery of Shapiro et al. will still work as originally intended. The only role of the plausibility values was to define a total pre-order on situations (Shapiro et al. 2011, p. 169 footnote), which we now get by comparing abnormalities.

Since abnormality predicates will define the plausibility of situations in a fixed, domain-independent way, to specify what an agent considers plausible for a particular domain we need to use the accessibility relation. To uniquely specify the accessibility relation we will use *only-knowing* (Lakemeyer and Levesque 1998) (only-knowing, unlike only-believing, does not assign plausibility levels). We can define an only-knowing operator `OKnows` as follows:

Definition 8 (Lakemeyer and Levesque (1998, p. 323))**.**

$$\mathtt{OKnows}(\phi, s) \stackrel{\text{def}}{=} (\forall s').\mathtt{SameHist}(s', s) \supset \mathtt{B}(s', s) \equiv \phi[s']$$

where $\mathtt{SameHist}(s', s)$ is the formula defined by Lakemeyer and Levesque that is true when s and s' have the same action histories from possibly different initial situations.

That is, ϕ is all that is known if, among all situations with the same action history, the accessible ones are exactly those in which ϕ is true. (Note that situations which do not have the same action history will not be accessible in any case.)

Finally, in order to use the only-knowing operator in the expected way, we need to ensure that our axioms entail that enough initial situations exist. For example, in order to get (for an action theory Σ) that $\Sigma \models \mathtt{OKnows}(P, \mathtt{S}_0) \supset \neg\mathtt{Bel}(\neg P, \mathtt{S}_0)$, i.e., that only-knowing P entails that $\neg P$ is disbelieved, we need Σ to entail that there exists an initial situation in which P is true. Similarly to Levesque, Pirri, and Reiter (1998, p. 173), we can include among the foundational axioms a second-order axiom that specifies there are initial situations with all combinations of fluent values. Then we can specify what an agent considers plausible by having them only-know a knowledge base that relates regular predicates to abnormality ones. To illustrate, by including the extra foundational axiom in an action theory Σ, we have that if all that is known is that non-abnormal birds fly and there is a bird, then it will be believed that the bird flies:

$$\Sigma \models \mathtt{OKnows}((\textsc{Bird} \wedge \neg\mathtt{Ab} \supset \textsc{Fly}) \wedge \textsc{Bird}), \mathtt{S}_0) \supset \mathtt{Bel}(\textsc{Fly}, \mathtt{S}_0).$$

So, our proposal for how to write an action theory Σ is as follows. Take an action theory Σ' of the sort that Shapiro et al. considered, and construct Σ by making these modifications to Σ': include an axiom of the form $\mathtt{OKnows}(\phi, \mathtt{S}_0)$ to specify the initial accessibility relation, redefine `MP` as in Definition 7, include the additional axiom for the existence of initial situations among the foundational axioms, and include the SSAs for the $\mathtt{Ab}_i$ predicates instead of for `pl`. We will call such an action theory Σ an *immutable abnormality action theory* (IAAT).

3.4 Examples

To demonstrate their features, we now apply IAATs by formalizing some simple examples.

Example 1 Our first example is meant to give an idea of how our approach works in general, while also illustrating that we can easily represent independent beliefs (avoiding the drowning problem). We revisit the problem from §2.4 about how to consider P and Q independently plausible.

In the domain for this problem, there are two fluents, $P(s)$ and $Q(s)$, whose values never change, and two sensing actions, SENSEP and SENSEQ, which respectively sense the values of P and Q. We'll also make use of two abnormality fluents, $\mathtt{Ab}_1(s)$ and $\mathtt{Ab}_2(s)$, of the same priority. In $\mathtt{S}_0$, the actual initial situation, P and Q are false. However, the agent does not know this. Instead, its knowledge base says that P is true (unless there is an abnormality) and Q is true (unless there is a different abnormality). For our action theory, we can axiomitize this description as follows:

$$P(\mathtt{do}(a,s)) \equiv P(s) \qquad Q(\mathtt{do}(a,s)) \equiv Q(s)$$
$$\mathtt{SF}(\text{SENSEP},s) \equiv P(s) \quad \mathtt{SF}(\text{SENSEQ},s) \equiv Q(s)$$
$$\neg P(\mathtt{S}_0) \wedge \neg Q(\mathtt{S}_0)$$
$$\mathtt{OKnows}((\neg \mathtt{Ab}_1 \supset P) \wedge (\neg \mathtt{Ab}_2 \supset Q), \mathtt{S}_0)$$

Initially, the accessible situations are exactly those initial situations where $(\neg \mathtt{Ab}_1 \supset P) \wedge (\neg \mathtt{Ab}_2 \supset Q)$ is true. Because belief is defined as what is true in the accessible situations with the fewest abnormalities, the agent initially (mistakenly) believes $P \wedge Q$. If it performs the sensing action SENSEP, it will come to correctly believe that P is false (but retain its belief that Q is true). If it then also performs SENSEQ, it will correctly believe that both P and Q are false. The proposition below formalizes these claims.

Proposition 2. *Let Σ be the IAAT described above. Then*

$$\Sigma \models \mathtt{Bel}(P \wedge Q, \mathtt{S}_0)$$
$$\Sigma \models \mathtt{Bel}(\neg P \wedge Q, \mathtt{do}(\text{SENSEP}, \mathtt{S}_0))$$
$$\Sigma \models \mathtt{Bel}(\neg P \wedge \neg Q, \mathtt{do}([\text{SENSEP}, \text{SENSEQ}], \mathtt{S}_0))$$

Example 2 This example will show the benefits of being able to define an unbounded number of plausibility levels.

Consider a language with the unary relational fluent CONSPIRATOR, where the intended meaning of $\text{CONSPIRATOR}(x)$ is that x is part of a conspiracy. There is one (sensing) action, $\text{REVEAL}(x)$, which reveals to the agent whether $\text{CONSPIRATOR}(x)$ is true. Who is a conspirator never changes, and in the actual initial situation $\mathtt{S}_0$, everyone is a conspirator. However, the agent thinks that situations with fewer conspirators are more plausible:

$$\mathtt{SF}(\text{REVEAL}(x), s) \equiv \text{CONSPIRATOR}(x, s)$$
$$\text{CONSPIRATOR}(x, \mathtt{do}(a, s)) \equiv \text{CONSPIRATOR}(x, s)$$
$$\text{CONSPIRATOR}(x, \mathtt{S}_0)$$
$$\mathtt{OKnows}\big((\forall x)\neg \mathtt{Ab}(x, \mathtt{now}) \supset \neg \text{CONSPIRATOR}(x, \mathtt{now}), \mathtt{S}_0\big)$$

The following proposition says that agent always believes that the only conspirators are those that have been revealed.

Proposition 3. *Let Σ be the IAAT described above, and let $\mathtt{C}_1, \mathtt{C}_2, \mathtt{C}_3, \ldots$ be constant symbols. Then for any k,*

$$\Sigma \models \mathtt{Bel}\Big((\forall x)\text{CONSPIRATOR}(x, \mathtt{now}) \equiv \Big(\bigvee\nolimits_{i=1}^{k} x = \mathtt{C}_i\Big),$$
$$\mathtt{do}([\text{REVEAL}(\mathtt{C}_1), \ldots, \text{REVEAL}(\mathtt{C}_k)], s)\Big)$$

Example 3 This example illustrates how we can use abnormalities with different priority levels.

The domain involves a light. There are two actions, the sensing action SENSELIT that senses whether the light is on (LIT), and the action FLIPUP, which flips the light switch up (UP) and also turns the light on (LIT) if it is not burnt out (BURNT). The agent knows that initially the light is on iff the switch is up and the light isn't burnt out (and the environment dynamics ensure this relationship continues to hold at all times). In the real initial situation, the switch is up but the light is burnt out. The agent initially considers that it would be implausible for the switch to be down and even more implausible for the light to be burnt out. In formalizing all this below, we make use of two abnormality predicates, $\mathtt{Ab}_1(s)$ and $\mathtt{Ab}_2(s)$, where $\mathtt{Ab}_2(s)$ has higher priority.

$$\text{BURNT}(\mathtt{do}(a,s)) \equiv \text{BURNT}(s)$$
$$\text{UP}(\mathtt{do}(a,s)) \equiv a = \text{FLIPUP} \vee \text{UP}(s)$$
$$\text{LIT}(\mathtt{do}(a,s)) \equiv (a = \text{FLIPUP} \wedge \neg\text{BURNT}(s)) \vee \text{LIT}(s)$$
$$\mathtt{SF}(\text{SENSELIT}) \equiv \text{LIT}(s) \qquad \mathtt{SF}(\text{FLIPUP}) \equiv \mathtt{True}$$
$$\neg\text{LIT}(\mathtt{S}_0) \wedge \text{UP}(\mathtt{S}_0) \wedge \text{BURNT}(\mathtt{S}_0)$$
$$\mathtt{OKnows}([\neg \mathtt{Ab}_1 \supset \text{UP}] \wedge [\neg \mathtt{Ab}_2 \supset \neg\text{BURNT}] \wedge$$
$$[(\text{UP} \wedge \neg\text{BURNT}) \equiv \text{LIT}], \mathtt{S}_0)$$

The agent will at first believe the light is on. After sensing that it isn't, the agent will then believe (incorrectly) that the switch is down. After also performing the FLIPUP action and sensing again, the agent will finally realize that the light is burnt out. This is formalized by the proposition below.

Proposition 4. *Let Σ be the IAAT described above. Then*

$$\Sigma \models \mathtt{Bel}(\text{LIT} \wedge \text{UP} \wedge \neg\text{BURNT}, \mathtt{S}_0)$$
$$\Sigma \models \mathtt{Bel}(\neg\text{LIT} \wedge \neg\text{UP} \wedge \neg\text{BURNT}, \mathtt{do}(\text{SENSELIT}, \mathtt{S}_0))$$
$$\Sigma \models \mathtt{Bel}(\neg\text{LIT} \wedge \text{UP} \wedge \text{BURNT},$$
$$\mathtt{do}([\text{SENSELIT}, \text{FLIPUP}, \text{SENSELIT}], \mathtt{S}_0))$$

4 Changing plausibility over time

Shapiro et al. specified that the plausibilities of situations never changed, and we have followed suit by keeping abnormalities fixed. An obvious alternative would be to instead allow actions to change what is abnormal. This could be useful for reasoning about *exogenous* actions, such as rain starting, or a flood occurring. Intuitively, the situation resulting from one of those actions could be more plausible than the other.

There is one thing to be careful with when updating plausibilities in this way. The agent believes what is true in all the *currently least abnormal* accessible situations, regardless of how many abnormalities previously existed. So if we write an action theory so as to say that an action removes or adds an abnormality, we have to be careful that what we mean is that the occurrence of that action really does make the situation (with its history) more or less plausible.

Shapiro and Pagnucco (2004) did generalize the framework of Shapiro et al. to allow exogenous actions, but in that work the agent could not compare the plausibility of exogenous actions, but just assumed there were as few exogenous actions in the past as possible. To be more precise, belief

was defined as truth in the "minimal" situations, where minimality was defined in terms of pl values (as in Shapiro et al.) except that ties in pl values were broken by favoring situations with shorter histories. We can generalize that.

Shapiro and Pagnucco divided actions into two types, exogenous and endogenous. They had unary predicates Exo and Endo to identify them. They required that exogenous actions not provide useful sensing information, by having the axiom $\texttt{Exo}(a) \supset (\forall s).\texttt{SF}(a, s)$. Furthermore, instead of the axioms constraining B that we have previously seen, they used an axiom that can be written as

$$(\forall s', s).\texttt{B}(s', s) \equiv \texttt{SameVisHist}(s, s'),$$

where $\texttt{SameVisHist}(s, s')$ is an abbreviation for a formula saying that s and s' have the same endogenous actions in their histories in the same order (and with the same sensing results), but with possibly different exogenous actions interleaved among them. Intuitively, this reflects how the agent is aware what it itself does, but is not aware of exogenous actions (except of what it can infer through sensing).

As Shapiro and Pagnucco note, this axiom does more than a successor state axiom usually does – it also describes B in initial situations. In their approach the accessibility relation is domain-independent, and it is only by specifying the plausibility function that the axiomatizer gets to determine what the agent believes. This is rather the opposite of the approach we have been taking, where the plausibility of an initial situation is fixed by what abnormalities exist there, and the beliefs of the agent are determined by the axiomatizer specifying the accessibility relation (with only-knowing).

Instead of using their axiom for B, we can specify what the agent knows was true in the initial situation by including a sentence of the form $\texttt{Oinit}(\phi)$ in an action theory, where

$$\texttt{Oinit}(\phi) \stackrel{\text{def}}{=} (\forall s', s).\texttt{B}(s', s) \equiv \big[\texttt{SameVisHist}(s, s') \wedge (\exists s^*).\texttt{Init}(s^*) \wedge s^* \sqsubseteq s' \wedge \phi[s^*]\big].$$

$\texttt{Oinit}(\phi)$ says that accessible situations must have the same endogenous actions in the same order, and furthermore the knowledge base ϕ must have been true at the initial situations in their histories. Note that this does not necessarily mean that the agent initially believes ϕ, since they may think that ϕ could have been made false by exogenous actions.

So, now we can consider *mutable abnormality action theories* (MAATs). MAATs are like IAATs, except that abnormality predicates are now allowed to have different SSAs, MAATs specify which actions are exogenous (and that those action don't provide sensing information), and MAATs use $\texttt{Oinit}(\phi)$ to specify B.

Example 4: counting exogenous actions First, let's consider how we might emulate the way Shapiro and Pagnucco counted exogenous actions to determine plausibility. We can define a fluent CLOCK that counts actions:

$$\textsc{Clock}(i, \texttt{do}(a, s)) \equiv (\exists j).i = j + 1 \wedge \textsc{Clock}(j, s)$$

We can then specify that $\texttt{Ab}(i, s)$ is true if there is a situation $s' \sqsubset s$ where $\textsc{Clock}(i, s')$ was true and in which an exogenous action occurred.

$$\texttt{Ab}(i, \texttt{do}(a, s)) \equiv (\textsc{Clock}(i, s) \wedge \texttt{Exo}(a)) \vee \texttt{Ab}(i, s)$$

By including $\texttt{Oinit}(\textsc{Clock}(0) \wedge (\forall i).\neg\texttt{Ab}(i) \wedge [i \neq 0 \supset \neg\textsc{Clock}(i)] \wedge \alpha)$ in the MAAT – where α is any formula, and the rest specifies that the agent knows the initial time was 0 and there were no abnormalities then – we then have that for an *accessible* situation s, $\texttt{Ab}(i, s)$ is true iff the ith action in the history of s was exogenous. Consider how this affects the plausibility of accessible situations. If all other abnormalities have higher priority than Ab and never change, this amounts to breaking ties in plausibility by counting exogenous actions, as in Shapiro and Pagnucco's approach.

Example 5: the plausibility of rain versus flooding This example, in which we will model rain as more plausible than flooding, shows how we can go beyond just counting exogenous actions to determine the plausibility of situations. We have two exogenous actions, rain (RAIN) and flooding (FLOOD) either of which causes the ground to be wet (WET). For the purposes of this example, rain and flooding will be modeled as occurring independently. There is an endogenous sensing action SEE which checks if the ground is wet.

$$\textsc{Wet}(\texttt{do}(a, s)) \equiv (a = \textsc{rain} \vee a = \textsc{flood}) \vee \textsc{Wet}(s)$$
$$\texttt{SF}(\textsc{see}, s) \equiv \textsc{Wet}(s)$$

We also have two abnormality fluents, $\texttt{Ab}_1$ and $\texttt{Ab}_2$, where $\texttt{Ab}_1$ has higher priority than $\texttt{Ab}_2$. Suppose we have an SSA for CLOCK as before. We can set up the SSAs for $\texttt{Ab}_1$ and $\texttt{Ab}_2$ so that flooding at time i causes $\texttt{Ab}_1(i)$ to become true, and rain at time i causes $\texttt{Ab}_2(i)$ to become true:

$$\texttt{Ab}_1(i, \texttt{do}(a, s)) \equiv [\textsc{Clock}(i, s) \wedge a = \textsc{flood}] \vee \texttt{Ab}_1(i, s)$$
$$\texttt{Ab}_2(i, \texttt{do}(a, s)) \equiv [\textsc{Clock}(i, s) \wedge a = \textsc{rain}] \vee \texttt{Ab}_2(i, s)$$

Furthermore, the agent thinks that initially the ground was not wet, the time was 0, and there were no abnormalities.

$$\texttt{Oinit}(\neg\textsc{Wet} \wedge \textsc{Clock}(0) \wedge (\forall i).\neg\texttt{Ab}_1(i) \wedge \neg\texttt{Ab}_2(i) \wedge [i \neq 0 \supset \neg\textsc{Clock}(i)])$$

The next proposition says that after an exogenous action occurs and the agent then senses that the ground is wet, the agent believes (possibly mistakenly) that it rained. The reason for this is that the agent knows that it either rained or flooded, but considers the rain more plausible.

Proposition 5. *Let Σ be the MAAT described above. Then*

$$\Sigma \models \texttt{Bel}((\exists s).\texttt{do}(\textsc{rain}, s) \sqsubset \texttt{now}, \texttt{do}([\textsc{rain}, \textsc{see}], \texttt{S}_0))$$
$$\Sigma \models \texttt{Bel}((\exists s).\texttt{do}(\textsc{rain}, s) \sqsubset \texttt{now}, \texttt{do}([\textsc{flood}, \textsc{see}], \texttt{S}_0))$$

Proof. In either $\texttt{do}([\textsc{rain}, \textsc{see}])$ or $\texttt{do}([\textsc{flood}, \textsc{see}], \texttt{S}_0)$, the accessible situations all have at least one RAIN or FLOOD in their history. The most plausible such situation has (just) one RAIN action. □

Example 6: the fate of abandoned money Sometimes, for an exogenous action to have occurred may seem more likely than not. For example, if there was money on the street, you might expect that it will have been taken.

Suppose that there is one exogenous action, STEAL (and possibly some number of endogenous actions). There is a fluent ONSTREET indicating that money is on the street.

The STEAL action results in any money on the street disappearing. For there to be money on the street is abnormal (Ab). The agent believes that initially there was money on the street (abnormally). This description is formalized below:

$$\begin{aligned}\textsc{OnStreet}(\texttt{do}(a,s)) &\equiv (\textsc{OnStreet}(s) \wedge a \neq \textsc{steal})\\ \texttt{Ab}(\texttt{do}(a,s)) &\equiv (\textsc{OnStreet}(s) \wedge a \neq \textsc{steal})\\ \texttt{0init}(&\textsc{OnStreet} \wedge \texttt{Ab})\end{aligned}$$

Recall that we are no longer assuming that an agent realizes when it is in an initial situation. An agent in $\mathtt{S}_0$ can believe (mistakenly) that some exogenous actions have taken place. In this example, although the agent believes in $\mathtt{S}_0$ that initially there was money on the street, it also believes in $\mathtt{S}_0$ that the money has already been stolen.

Proposition 6. *Let Σ be the MAAT described above. Then*

$$\Sigma \models \texttt{Bel}((\exists s).\texttt{do}(\textsc{steal}, s) = \texttt{now}, \mathtt{S}_0).$$

Proof. The initially accessible situations are initial situations (where Ab is true) and situations where STEAL just occurred (and Ab is false). The latter are more plausible. □

5 Alternatives to CBC

CBC has been seldom used in the literature. It does have limitations; for example, in contrast to regular circumscription, CBC requires the axiomatizer to make the stronger commitment that any set of $n+1$ abnormalities is less plausible than any set of n abnormalities (if all are at the same priority level). Below, we provide support for why CBC is an appropriate choice for specifying plausibility levels by considering some alternatives. First, we show how CBC is more general than another technique that might be considered, lexicographic entailment. Then, we explain why we could not have used regular circumscription in the way we have used CBC.

5.1 Lexicographic entailment

Recall that Schwering and Lakemeyer's only-believing operator determined a plausibility ordering like that given by System Z. There is no reason that we can't define versions of only-believing based on other systems from the extensive literature on using conditionals for default reasoning (Geffner and Pearl 1992; Goldszmidt, Morris, and Pearl 1993; Benferhat et al. 1993; Lehmann 1995; Kern-Isberner and Eichhorn 2014; Beierle et al. 2017). In this section we will consider using one of these systems, lexicographic entailment, in defining an alternative "only-believing" operator. We will then show that CBC is a more general approach.

Lexicographic entailment comes from the work of Benferhat et al. (1993) and Lehmann (1995). The version of lexicographic entailment we'll describe is based on the presentation by Eiter and Lukasiewicz (2000) of *lex*$_{\text{p}}$-entailment.

In this system, a knowledge base is given as a pair $\langle \alpha, \Gamma \rangle$ where α is a sentence and $\Gamma = \{\phi_1 \Rightarrow \psi_1, \ldots, \phi_m \Rightarrow \psi_m\}$ is a set of conditionals (again, '$\Rightarrow$' is not the material conditional). Traditionally, α, ϕ_i, and ψ_i were considered to be propositional, but we can let them be first-order. Each conditional $\phi_i \Rightarrow \psi_i$ is associated with a *priority* level from $\{1, \ldots, k\}$ (where 1 is the most important). Given $\langle \alpha, \Gamma \rangle$, we can associate with every interpretation $\mathfrak{I}$ a *preference vector* $\vec{\ell}(\mathfrak{I}) \in \{0, \ldots, m\}^k$, where the ith entry of $\vec{\ell}(\mathfrak{I})$ is the number of values of j for which $(\phi_j \Rightarrow \psi_j)$ is a priority i conditional and $\mathfrak{I} \not\models (\phi_j \supset \psi_j)$.

We will say that $\langle \alpha, \Gamma \rangle$ lexicographically entails $\phi \Rightarrow \psi$, written $\langle \alpha, \Gamma \rangle \models_{\mathsf{lex}} \phi \Rightarrow \psi$, if ψ is true in every interpretation $\mathfrak{I}$ with minimal $\vec{\ell}(\mathfrak{I})$ such that $\mathfrak{I} \models \alpha \wedge \phi$. As with abnormality vectors in CBC, minimality is determined by lexicographic comparison: $\vec{\ell}(\mathfrak{I}_1) < \vec{\ell}(\mathfrak{I}_2)$ if there exists an i so that $\vec{\ell}(\mathfrak{I}_1)_i < \vec{\ell}(\mathfrak{I}_2)_i$ and for all $j < i$ we have $\vec{\ell}(\mathfrak{I}_1)_j \leq \vec{\ell}(\mathfrak{I}_2)_j$.

Note that there are only a finite number of distinct vectors in the image of $\vec{\ell}(\cdot)$, so we can number them $\vec{\ell}_1, \vec{\ell}_2, \ldots \vec{\ell}_N$ so that $\vec{\ell}_i < \vec{\ell}_{i+1}$. We could define another only-believing operator, which we'll call $\mathbf{O}_{\mathsf{lex}}$, by "embedding" lexicographic entailment within it. For $e = (e_1, e_2, \ldots)$ an epistemic state, we define $e \models \mathbf{O}_{\mathsf{lex}}(\Gamma)$ to hold iff each e_i contains every world w where $\vec{\ell}(w) \leq \vec{\ell}_i$ (let $e_i = e_N$ when $i \geq N$).

This new form of only-believing avoids the drowning problem, insofar as lexicographic entailment does. For example, we have the following:

Proposition 7.

$$\mathbf{O}_{\mathsf{lex}}(\texttt{True} \Rightarrow P, \texttt{True} \Rightarrow Q) \models \mathbf{B}(\neg P \Rightarrow Q) \wedge \mathbf{B}(\neg Q \Rightarrow P)$$

Proof. In the epistemic state $e = (e_1, e_2, \ldots)$ that satisfies the left-hand side, e_1 contains the worlds where both $(\texttt{True} \supset P)$ and $(\texttt{True} \supset Q)$ are true, and e_2 contains the worlds where at least one of those conditionals is true. So Q is true at the most plausible $\neg P$-worlds (which are in e_2), and similarly P is true at the most plausible $\neg Q$-worlds. □

As their similarity suggests, there is a sense in which lexicographic entailment can be easily translated into CBC.

Lemma 1. *Suppose that $\mathtt{Ab}_1, \ldots \mathtt{Ab}_m$ are all the abnormality predicates and are all 0-ary, and $\phi_1, \ldots \phi_m, \psi_i, \ldots, \psi_m$ are sentences not including any $\mathtt{Ab}_i$ symbol. Let us define $\vec{\ell}(\mathfrak{I})$ relative to $\langle \alpha, \{\phi_1 \Rightarrow \psi_1, \ldots, \phi_m \Rightarrow \psi_m\}\rangle$, where the priority of $\psi_i \Rightarrow \phi_i$ is the same as the priority of $\mathtt{Ab}_i$. Then for every interpretation $\mathfrak{I}$ such that $\mathfrak{I} \models \bigwedge\{\neg\mathtt{Ab}_i \equiv (\phi_i \supset \psi_i) : 1 \leq i \leq m\}$, we have $\vec{c}(\mathfrak{I}) = \vec{\ell}(\mathfrak{I})$.*

Proof. If $\mathfrak{I} \models \bigwedge\{\neg\mathtt{Ab}_i \equiv (\phi_i \supset \psi_i) : 1 \leq i \leq m\}$, then for each i such that $\mathfrak{I} \not\models (\phi_i \supset \psi_i)$, we have $\mathfrak{I} \models \mathtt{Ab}_i$ (and vice versa). The definitions of the abnormality vector $\vec{c}(\mathfrak{I})$ and preference vector $\vec{\ell}(\mathfrak{I})$ make them the same in that case. □

Proposition 8. *Let $\mathtt{Ab}_1, \ldots, \mathtt{Ab}_m$, $\phi_1, \ldots \phi_m$, $\psi_i, \ldots, \psi_m$, and $\vec{\ell}(\mathfrak{I})$ be as in Lemma 1 above. Suppose that α, β_1, and β_2 are sentences not including any abnormality symbols. Then $\langle \alpha, \{\phi_1 \Rightarrow \psi_1, \ldots, \phi_m \Rightarrow \psi_m\}\rangle \models_{\mathsf{lex}} \beta_1 \Rightarrow \beta_2$ iff $\{\alpha \wedge \beta_1\} \cup \{\neg\mathtt{Ab}_i \equiv (\phi_i \supset \psi_i) : 1 \leq i \leq m\} \models_{\mathsf{card}} \beta_2$.*

Proof. Immediate from Lemma 1. □

This resembles how *formula circumscription* (McCarthy 1986) can be defined in terms of (traditional) *predicate circumscription*. Lexicographic entailment is essentially a form of cardinality-based formula circumscription.

This relationship between CBC and lexicographic entailment is straightforward but to the best of our knowledge has not been previously reported on. Furthermore, CBC works sensibly in the first-order case; it's easy to get an infinite number of distinct abnormality vectors (by having the abnormality predicates take arguments), as we saw in Example 2 in §3.4. On the other hand, $\mathbf{O}_{\mathsf{lex}}$ only gives us at most $m+1$ distinct plausibility levels. We should however note that there is a first-order version of lexicographic entailment from Benferhat and Baida (2004), which is similar to CBC, though defined in a more complicated way (it involves considering what is entailed by "weakened" knowledge bases in which universally quantified formulas have been syntactically modified by listing exceptions to them).

5.2 Other forms of circumscription

The framework of Shapiro et al. obeys (a slightly modified version) of the AGM postulates for belief revision (Alchourrón, Gärdenfors, and Makinson 1985). This remains true when using CBC to describe the plausibility levels instead of the `pl` function. However, if we tried to make use of regular circumscription instead of CBC, that would cause problems, as we explain in this section.

Following Shapiro et al. (2011, Definitions 30–32), we define the belief state $K(t)$ of an agent in situation t, the expansion $t+\phi$, and the revision $t*\phi$ (all relative to a model $\mathfrak{I}$ of the action theory Σ) as follows: $K(t) \stackrel{\text{def}}{=} \{\varphi : \mathfrak{I} \models \mathtt{Bel}(\varphi, t)\}$, $t+\phi \stackrel{\text{def}}{=} \{\varphi : \mathfrak{I} \models \mathtt{Bel}(\phi \supset \varphi, t)\}$, and $t*\phi \stackrel{\text{def}}{=} \mathtt{do}(A_\phi, t)$, where A_ϕ is a revision action for the formula ϕ (which means the agent will no longer consider $\neg\phi$-situations possible, provided $\mathfrak{I} \models \phi$).

Shapiro et al.'s translation of the AGM axioms into this notation included the following: If $\neg\phi \notin K(t)$, then $t+\phi \subseteq K(t*\phi)$. That is, if an agent believes a material conditional and doesn't disbelieve its antecedent, then after revising by the antecedent the agent should believe the consequent.

We will show that this axiom can be violated if regular circumscription is used. Suppose that Σ is an action theory like the IAATs we considered before, except the comparison of abnormality predicates by cardinality is replaced by subset inclusion. Suppose that $\Sigma \models \mathtt{OKnows}(\mathtt{Ab}_1 \vee \mathtt{Ab}_2, t)$ and consider a model $\mathfrak{I}$ of Σ such that $\mathfrak{I} \models \phi_0[t]$, where ϕ_0 stands for $\neg(\mathtt{Ab}_1 \wedge \neg\mathtt{Ab}_3)$. Observe that $\neg\mathtt{Ab}_2 \notin K(t)$. Also, note that, we have $\mathfrak{I} \models \mathtt{Bel}(\phi_0 \supset \mathtt{Ab}_2, t)$, so $(\phi_0 \supset \mathtt{Ab}_2) \in K(t)$, and so $\mathtt{Ab}_2 \in t+\phi_0$. However, we also have that $\mathtt{Ab}_2 \notin K(t*\phi_0)$, contradicting the AGM axiom.

So regular circumscription cannot be used in the way we have used CBC. For the same reason, we also could not use an alternative form of CBC that – instead of summing together the cardinalities of the extensions of all predicates of the same priority – compared cardinalities for each predicate individually (see Moinard (2000, Remark 14)).

6 Related work

Here we briefly discuss a few as-yet-unmentioned works.

Pagnucco et al. (2013), working within the framework of Shapiro et al., for the purposes of implementation (for the use by a robot in interpreting directions) suggested a way of constraining the initial plausibility levels which resembles our approach. The idea is that a number of literals referring to the initial situation are "told" to the robot, and initial situations where more of those literals are true (taking into account priorities given to the fluent symbols in the literals) are constrained to be more plausible. Pagnucco et al. do not discuss using the accessibility relation to associate more complex sentences with these "told" literals, in contrast to the way we use only-knowing to associate abnormalities with other things.

The approach of del Val and Shoham (1994) to belief revision and update in a variant of the situation calculus also (like ours) featured abnormality predicates. However, their use of circumscription was to minimize change of "persistent" properties from situation to situation (they did not have Reiter-style successor state axioms).

Fang and Liu (2013) considered belief change in a multi-agent version of the situation calculus, which could also model actions that an agent was unaware of (like our exogenous actions). Following work in dynamic epistemic logic (Baltag and Smets 2008), they made use of two plausibility orderings, one on situations and one on actions, and updated the plausibility of situations by giving priority to the plausibility of the last action to have been performed (the so-called "action-priority update"). This is in the spirit of the importance placed on recent information in the AGM approach, but we would argue that is not the most natural way to reason about exogenous actions.

7 Conclusion

In order to apply Shapiro et al.'s framework for iterated belief change to any practical problem – from in robotics, to understanding stories (Klassen, Levesque, and McIlraith 2017) – it is necessary to first specify the plausibility levels of accessible situations somehow. We have presented a way of using cardinality-based circumscription (CBC) for this. We have shown how this way avoids the drowning problem and the limitation to finitely many plausibility levels that are issues with Schwering and Lakemeyer's approach. Furthermore, by allowing abnormality fluents to change over time, we have developed an approach to handling exogenous actions which is more general than Shapiro and Pagnucco's, in that we can (for example) associate different plausibilities with different actions, and even make the *non*-occurrence of exogenous actions implausible.

We have also shown how to describe the form of CBC from Klassen, Levesque, and McIlraith (2017) in second-order logic. We formally characterized the close relationship between CBC and lexicographic entailment. For future work, we think that other areas of interest in reasoning about action, such as noisy sensing, non-determinism, and failed actions probably could be given useful treatments in terms of abnormalities as well.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Alchourrón, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic* 50(2):510–530.

Baltag, A., and Smets, S. 2008. A qualitative theory of dynamic interactive belief revision. In *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, Texts in Logic and Games 3. Amsterdam University Press. 11–58.

Beierle, C.; Falke, T.; Kutsch, S.; and Kern-Isberner, G. 2017. System ZFO: Default reasoning with system Z-like ranking functions for unary first-order conditional knowledge bases. *International Journal of Approximate Reasoning* 90:120–143.

Benferhat, S., and Baida, R. E. 2004. A stratified first order logic approach for access control. *International Journal of Intelligent Systems* 19(9):817–836.

Benferhat, S.; Cayrol, C.; Dubois, D.; Lang, J.; and Prade, H. 1993. Inconsistency management and prioritized syntax-based entailment. In *IJCAI 1993*, 640–645.

Brachman, R. J., and Levesque, H. J. 2004. *Knowledge Representation and Reasoning*. Morgan Kaufmann.

Darwiche, A., and Pearl, J. 1997. On the logic of iterated belief revision. *Artificial Intelligence* 89(1):1–29.

del Val, A., and Shoham, Y. 1994. A unified view of belief revision and update. *Journal of Logic and Computation* 4(5):797–810.

Demolombe, R., and Parra, P. P. 2006. Belief revision in the situation calculus without plausibility levels. In *ISMIS 2006*, 504–513.

Eiter, T., and Lukasiewicz, T. 2000. Default reasoning from conditional knowledge bases: Complexity and tractable cases. *Artificial Intelligence* 124(2):169–241.

Fang, L., and Liu, Y. 2013. Multiagent knowledge and belief change in the situation calculus. In *AAAI 2013*, 304–312.

Friedman, N., and Halpern, J. Y. 1999. Modeling belief in dynamic systems, part II: Revision and update. *Journal of Artificial Intelligence Research (JAIR)* 10:117–167.

Geffner, H., and Pearl, J. 1992. Conditional entailment: Bridging two approaches to default reasoning. *Artificial Intelligence* 53(2):209–244.

Goldszmidt, M.; Morris, P.; and Pearl, J. 1993. A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(3):220–232.

Katsuno, H., and Mendelzon, A. O. 1991. Propositional knowledge base revision and minimal change. *Artificial Intelligence* 52(3):263–294.

Kern-Isberner, G., and Eichhorn, C. 2014. Structural inference from conditional knowledge bases. *Studia Logica* 102(4):751–769.

Klassen, T. Q.; Levesque, H. J.; and McIlraith, S. A. 2017. Towards representing what readers of fiction believe. In *COMMONSENSE 2017*.

Lakemeyer, G., and Levesque, H. J. 1998. AOL: a logic of acting, sensing, knowing, and only knowing. In *KR 1998*, 316–327.

Lehmann, D. 1995. Another perspective on default reasoning. *Annals of Mathematics and Artificial Intelligence* 15(1):61–82.

Levesque, H.; Pirri, F.; and Reiter, R. 1998. Foundations for a calculus of situations. *Electronic Transactions of AI (ETAI)* 2(3–4):159–178.

Liberatore, P., and Schaerf, M. 1995. Relating belief revision and circumscription. In *IJCAI 1995*, 1557–1566.

Liberatore, P., and Schaerf, M. 1997. Reducing belief revision to circumscription (and vice versa). *Artificial Intelligence* 93(1):261–296.

Lifschitz, V. 1994. Circumscription. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3. Oxford University Press. 297–352.

McCarthy, J., and Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4*, 463–502. Edinburgh University Press.

McCarthy, J. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence* 13(12):27–39.

McCarthy, J. 1986. Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence* 28(1):89–116.

Moinard, Y. 2000. Note about cardinality-based circumscription. *Artificial Intelligence* 119(1):259 – 273.

Pagnucco, M.; Rajaratnam, D.; Strass, H.; and Thielscher, M. 2013. Implementing belief change in the situation calculus and an application. In *LPNMR 2013*, 439–451.

Pearl, J. 1990. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In *TARK 1990*, 121–135.

Peppas, P. 2008. Chapter 8: Belief revision. In van Harmelen, F.; Lifschitz, V.; and Porter, B., eds., *Handbook of Knowledge Representation*, volume 3. Elsevier. 317–359.

Reiter, R. 2001. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT Press.

Scherl, R. B., and Levesque, H. J. 2003. Knowledge, action, and the frame problem. *Artificial Intelligence* 144(1):1 – 39.

Schwering, C., and Lakemeyer, G. 2014. A semantic account of iterated belief revision in the situation calculus. In *ECAI 2014*, 801–806.

Schwering, C.; Lakemeyer, G.; and Pagnucco, M. 2017. Belief revision and projection in the epistemic situation calculus. *Artificial Intelligence* 251:62–97.

Schwering, C. 2016. *Conditional Beliefs in Action*. Ph.D. Dissertation, RWTH Aachen University.

Shapiro, S., and Pagnucco, M. 2004. Iterated belief change and exogeneous actions in the situation calculus. In *ECAI 2004*, 878–882.

Shapiro, S.; Pagnucco, M.; Lespérance, Y.; and Levesque, H. J. 2011. Iterated belief change in the situation calculus. *Artificial Intelligence* 175(1):165–192.

Sharma, N., and Colomb, R. 1997. Towards an integrated characterisation of model-based diagnosis and configuration through circumscription policies. Technical Report 364, Department of Computer Science, University of Queensland.

Strasser, C., and Antonelli, G. A. 2016. Non-monotonic logic. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Stanford University, Winter 2016 edition. `https://plato.stanford.edu/archives/win2016/entries/logic-nonmonotonic/`.