Causality and the Qualification Problem

In L. C. Aiello and J. Doyle and S. Shapiro, editors, Proc. of the International Conference on Principles of Knowledge Representation and Reasoning, pages 51–62, Cambridge, MA, 1996. Morgan Kaufmann

Michael Thielscher*
International Computer Science Institute
1947 Center Street
Berkeley, CA 94704-1198

Abstract

In formal theories for reasoning about actions, the qualification problem denotes the problem to account for the many conditions which, albeit being unlikely to occur, may prevent the successful execution of an action. By a simple counter-example in the spirit of the well-known Yale Shooting scenario, we show that the common straightforward approach of globally minimizing such *abnormal disqualifications* is inadequate as it lacks an appropriate notion of causality. To overcome this difficulty, we propose to incorporate causality by treating the proposition that an action is qualified as a *fluent* which is initially assumed away by default but otherwise potentially indirectly affected by the execution of actions. Our formal account of the qualification problem includes the proliferation of explanations for surprising disqualifications and also accommodates so-called miraculous disqualifications. We moreover sketch a version of the fluent calculus which involves default rules to address abnormal disqualifications of actions, and which is provably correct wrt. our formal characterization of the qualification problem.

1 INTRODUCTION

A fundamental requirement for autonomous intelligent agents is the ability to reason about causality, which enables the agent to understand the world to an extent sufficient for acting intelligently on the basis of his or her knowledge as to the effects of actions. The *qualification problem* [McCarthy, 1977] in formal theories for reasoning about actions arises from the fact that generally the successful execution of actions depends on many more conditions than we are usually aware of. The reason for this unawareness is that most conditions are so likely to be satisfied that they are assumed away in case there is no evidence to the contrary.

A standard example to illustrate this is when we intend to start our car's engine, then we usually do not make sure that no potato in the tail pipe prevents us from doing so, despite the fact that a clogged tail pipe necessarily renders this action impossible.[1] While this *prima facie* ignorance is rational as it is generally impossible to verify all possible preconditions,[2] they cannot be completely disregarded in a sound formal model. Yet a proposition like "there is no potato in the tail pipe" should not be treated as a *strict* precondition in the formal specification of the action "start the engine" lest the reasoning agent always has to verify this condition before assuming that the action can be successfully executed. Moreover, it is often difficult if not impossible to even think of all conceivable disqualifications in advance [McCarthy, 1977].

Allowing to assume away all so-called *abnormal* disqualifications by default naturally implies that if further knowledge hints at any such unexpected disqualification, then the previous conclusion that the action in question be qualified needs to be withdrawn. Thus the entire process is intrinsically nonmonotonic. As a consequence, McCarthy's proposal was to employ circumscription with the aim of minimizing abnormal disqualifications [McCarthy, 1977; McCarthy, 1980; McCarthy, 1986]. Little has been achieved since then towards formally integrating this concept into a specific action formalism, or towards an assessment of its range of applicability. In fact, a surprisingly simple example illustrates that the straightforward global minimization of abnormal disqualifications is inadequate. The example shows some similarities to the problem—

*On leave from FG Intellektik, TH Darmstadt.

[1] According to [Ginsberg and Smith, 1988b], this example is due to McCarthy.

[2] Aside from the fact that besides a clear tail pipe there are lots of other disqualifying, albeit unlikely, obstacles, how can we ensure that after checking the tail pipe it does not become clogged during us walking to the front door and taking a seat, prior to trying to start the engine?

first illustrated with the Yale Shooting domain [Hanks and McDermott, 1987]—which occurs when neglecting causality in tackling the frame problem.

Imagine the following scenario: We can put a potato into the tail pipe whenever no abnormal disqualification prevents us from doing so (e.g., the potato surprisingly turns out to be too heavy); likewise we can start the engine except in case of an abnormal disqualification (like a potato in the tail pipe). Now, what would we predict as to the outcome of first trying to place a potato in the tail pipe and, then, trying to start the engine? Clearly, since nothing hints at an abnormal disqualification of the first action, we should expect this one to be successful. Then its effect (viz. a potato in the tail pipe) implies that the second action will be unqualified.

But what happens if abnormal disqualifications are globally minimized in this scenario? One minimal model is obviously obtained by considering the *put-potato* action qualified and the *start-engine* action unqualified, as expected. However, if instead the first action is assumed unqualified, then this in turn avoids the necessity of assuming a disqualification of the second. For if *put-potato* is not qualified, then it fails to produce what otherwise causes the disqualification of *start-engine*. Hence, in so doing we can construct a second minimal model for our scenario—which is clearly unintended.

The reason for the existence of the second, counter-intuitive model is that global minimization does not allow to distinguish disqualifications which can be explained from the standpoint of causality. Successfully introducing a potato into the tail pipe produces an effect which *causes* the fact that the second action, starting the engine, is unqualified. That is to say, while an abnormal disqualification of *put-potato* comes out of the blue in the unintended minimal model, an abnormal disqualification of *start-engine*, as claimed in the first minimal model, is easily explicable. One even tends to not call abnormal this situation since being unable to start the engine after having clogged the tail pipe is, after all, what one would normally expect. The reader might notice the similarities to the Yale Shooting problem: A gun that becomes magically unloaded while waiting deserves being called abnormal, whereas causality explains the death of the turkey if being shot at with a loaded gun [Hanks and McDermott, 1987].

The only existing alternative to global minimization of abnormalities as an approach to the qualification problem is based on *chronological ignorance* [Shoham, 1987; Shoham, 1988]. The basic idea there is to assume away by default abnormal, disqualifying circumstances, and simultaneously to prefer minimization of abnormalities at earlier timepoints. While this method treats our example scenario correctly, it is inherently incapable of handling non-deterministic actions, or non-deterministic information in general, as has already been argued elsewhere. A detailed account of this approach is given in the concluding discussion, Section 5.

Given the inadequacy of global minimization and the limited expressiveness of chronological ignorance, we propose a formal account of the qualification problem which incorporates a suitable concept of causality. We accomplish this by taking the proposition that an action is abnormally disqualified as a fluent, i.e., a proposition that may change its truth value in the course of time.[3] This proposition is assumed false, by default, *initially*, and by virtue of being fluent, it may be affected by the execution of an action and otherwise is subject to the general law of persistence. This helps to distinguish action disqualifications which are (indirectly) caused by actions that have been observed. As this method requires an appropriate treatment of indirect effects, we will adopt the approach to the *ramification problem* proposed in [Thielscher, 1997], where indirect effects are obtained according to so-called causal relationships among fluents. As a side gain, this enables us to account for *implicit* strict preconditions of actions, which are not part of an action specification but derive from certain domain constraints. This is sometimes considered part of the qualification problem, e.g. in [Ginsberg and Smith, 1988b; Lin and Reiter, 1994].

Aside from providing means to assume away abnormal disqualifications by default while properly taking into account possible causes for these disqualifications, the successful treatment of the qualification problem should include the proliferation of explanations in case an action has been—unexpectedly—observed unqualified. It may of course happen, though, that we are still unable to perform an action even if we have explicitly excluded, to the best of our knowledge, any imaginable preventing cause. However surprising this might be, it just shows us that we have only partial knowledge of the world. We call *miraculous* a disqualification which cannot be explained even if abnormal circumstances are granted. Consequently, miraculous disqualifications are to be minimized with higher priority than abnormal disqualifications which admit an explanation. Another characteristics of miraculous disqualifications is that they may occur or vanish even if, from our perspective, the situation has not changed. Again this is due to our lack of omniscience. The formal account of the qualification problem presented in this paper addresses both finding explanations for unexpectedly observed disqualifications and accounting for miraculous disqualifications.

We moreover sketch, on the basis of the *fluent calculus* [Hölldobler and Schneeberger, 1990; Thielscher,

[3] Throughout the paper, by "(dis-)qualified" we mean "physically (im-)possible." The refinement that actions may be unqualified *as to producing a certain effect* will be discussed at the end, in Section 5.

1997], an action calculus which includes a proper treatment of abnormal disqualifications. Since the qualification problem requires some sort of nonmonotonic feature, we employ *default rules* in the sense of [Reiter, 1980] to formalize the initial normality assumptions as well as the assumption that miraculous disqualifications do not occur. The resulting action calculus is provably correct wrt. our formal characterization of the qualification problem.

2 ACTIONS AND RAMIFICATIONS

2.1 A BASIC THEORY OF ACTIONS

The basic entities of action scenarios are *states*, each of which is a snapshot of the underlying dynamic system, i.e., the part of the world being modeled, at a particular instant. Formally, a state is determined by an assignment of truth values to a fixed set of propositional constants.[4]

Definition 1 Let $\mathcal{F}$ be a finite set of symbols called *fluent names*. A *fluent literal* is either a fluent name $f \in \mathcal{F}$ or its negation, denoted by $\overline{f}$. A set of fluent literals is *inconsistent* iff it contains some $f \in \mathcal{F}$ along with $\overline{f}$. A *state* is a maximal consistent set of fluent literals. ■

Notice that formally any combination of truth values denotes a state, which, however, might be considered impossible due to specific dependencies among the fluents (see below). Throughout the paper we assume the following notational conventions: If ℓ is a fluent literal, then $|\ell|$ denotes its affirmative component, that is, $|f| = |\overline{f}| = f$ where $f \in \mathcal{F}$. This notation extends to sets of fluent literals S as follows: $|S| = \{|\ell| : \ell \in S\}$. E.g., for each state S we have $|S| = \mathcal{F}$. Furthermore, if $\ell = \overline{f}$ is a negative fluent literal then $\overline{\ell}$ should be interpreted as f.

The elements of an underlying set of fluent names can be considered atoms for constructing (propositional) formulas to allow for statements about states. Each fluent literal and $\top$ (*tautology*) and $\bot$ (*contradiction*) are *fluent formulas*, and if F and G are fluent formulas then so are $F \wedge G$, $F \vee G$, $F \supset G$, and $F \equiv G$.[5] The notion of fluent formulas being *true* in a state S is based on defining a literal ℓ to be true if and only if $\ell \in S$. Fluent formulas provide means to distinguish states that cannot occur due to specific dependencies among particular fluents. Formulas which have to be satisfied in all states that are possible in a domain are also called *domain constraints*.

[4] The calculus described in Section 4 employs a more expressive language, which involves non-propositional fluents.

[5] As negation can be expressed through negative literals, we omit the standard connective " $\neg$ ". This is just for the sake of readability as it avoids too many different forms of negation.

Example 1 A basic version of the Potato In Tail Pipe scenario shall be formalized with the fluent names $\mathcal{F} = \{pot, clog, runs, heavy\}$ to state whether, respectively, there is a potato in the tail pipe, the tail pipe is clogged, the engine is running, and the potato is too heavy. The fluent formula

$$pot \supset clog \tag{1}$$

then expresses the fact that the tail pipe is clogged whenever it houses a potato. Taken as domain constraint, this formula is true, for instance, in the state $\{\overline{pot}, \overline{clog}, \overline{runs}, heavy\}$. ■

The second basic entity in theories of actions are the *actions* themselves, whose execution causes state transitions. Since stress shall lie on the qualification problem rather than on sophisticated methods of specifying the direct effects of actions, we employ a suitably simple, STRIPS-style [Fikes and Nilsson, 1971] notion of action specification. Each *action law* consists of

- A *condition* C, which is a set of fluent literals all of which must be contained in the state at hand in order to apply the action law.
- A (direct) *effect* E, which is a set of fluent literals, too, all of which hold in the resulting state after having applied the action law.

It is assumed that $|C| = |E|$, that is, condition and effect refer to the very same set of fluent names. This is just for the sake of simplicity, for it enables us to obtain the state resulting from the direct effect by simply removing set C from the state at hand and adding set E to it. This assumption does not impose a restriction of expressiveness since we allow several laws for a single action, and since any (unrestricted) action law can be replaced by an equivalent set of action laws which obey the assumption.

Definition 2 Let $\mathcal{F}$ be a set of fluent names, and let $\mathcal{A}$ be a finite set of symbols, called *action names*, such that $\mathcal{F} \cap \mathcal{A} = \{\}$. An *action law* is a triple $\langle C, a, E\rangle$ where C and E are consistent sets of fluent literals such that $|C| = |E|$, and $a \in \mathcal{A}$.

If S is a state, then an action law $\alpha = \langle C, a, E\rangle$ is *applicable* in S iff $C \subseteq S$. The *application* of α to S yields the state $(S \setminus C) \cup E$. ■

Obviously, S being a state, C and E being consistent, and $|C| = |E|$ guarantee $(S \setminus C) \cup E$ to be a state again—not necessarily, however, one which satisfies the underlying domain constraints.

Example 1 (continued) We define the action names *start* (starting the engine) and *put-p* (putting a potato into the tail pipe), which are accompanied by these action laws:

$$\begin{array}{c} \langle\, \{\overline{runs}\}, start, \{runs\}\,\rangle \\ \langle\, \{\overline{pot}\}, put\text{-}p, \{pot\}\,\rangle \end{array} \tag{2}$$

In words, starting the engine is possible if it is not running and causes it to do so; similarly, a potato may be added to the tail pipe. The second law, say, is applicable in the state $S = \{\overline{pot}, \overline{clog}, \overline{runs}, \overline{heavy}\}$ since $\{\overline{pot}\} \subseteq S$. Its application yields $(S \setminus \{\overline{pot}\}) \cup \{pot\}$, i.e., $\{pot, \overline{clog}, \overline{runs}, \overline{heavy}\}$, which constitutes a state but does not satisfy our constraint, $pot \supset clog$. ■

The example illustrates that a state obtained through the application of an action law may violate the underlying domain constraints since only direct effects have been specified: Putting a potato into the tail pipe has the *indirect* effect that the latter becomes clogged. The problem of accommodating additional, indirect effects is commonly referred to as the *ramification problem* [Ginsberg and Smith, 1988a]. Prior to discussing a suitable solution, observe that according to Definition 2 it is possible to construct a set of action laws which, given a state, contains more than one applicable law for a single action name. This can be used to formalize non-deterministic actions.

Example 2 Suppose we park our car in a neighborhood that is known for its suffering from a tail pipe marauder.[6] We therefore must expect that after waiting for a certain amount of time, a potato may have randomly been introduced into our car's tail pipe. This is formally captured by giving a non-deterministic specification of an action with the name *wait*. Let $\mathcal{F} = \{pot, clog, runs\}$ and $\mathcal{A} = \{wait, start\}$. Performing a *wait* action either has no effect at all, or else it causes *pot* become true provided there is not already a potato in the tail pipe. Accordingly, we employ the following two action laws:

$$\langle\, \{\}, wait, \{\}\,\rangle \quad \text{and} \quad \langle\, \{\overline{pot}\}, wait, \{pot\}\,\rangle \qquad (3)$$

Both of them are applicable, for instance, in the state $\{\overline{pot}, \overline{clog}, \overline{runs}\}$, which suggests two possible outcomes, viz. $\{\overline{pot}, \overline{clog}, \overline{runs}\}$ and $\{pot, \overline{clog}, \overline{runs}\}$. ■

2.2 THE RAMIFICATION PROBLEM

In [Thielscher, 1997] we propose to address the ramification problem by regarding the collection of fluent literals resulting from the computation of the *direct* effects merely as an intermediate state, which requires additional computation accounting for possible *indirect* effects. More specifically, a single indirect effect is obtained according to a directed *causal* relation between two particular fluents.

Definition 3 Let $\mathcal{F}$ be a set of fluent names. A *causal relationship* is an expression of the form ε `causes` ϱ `if` Φ where Φ is a fluent formula and ε and ϱ are fluent literals. ■

[6]This example has been suggested by Erik Sandewall (personal communication).

The intended reading is the following: Under condition Φ, the (previously obtained, direct or indirect) effect ε triggers the indirect effect ϱ. E.g., the causal relationship *pot* `causes` *clog* `if` $\top$ will be used below to state that the effect *pot* always gives rise to the additional effect *clog*. Causal relationships operate on pairs (S, E), where S denotes the current state and E contains all direct and indirect effects computed so far:

Definition 4 Let (S, E) be a pair consisting of a state S and a set of fluent literals E, then a causal relationship ε `causes` ϱ `if` Φ is *applicable* to (S, E) iff $\Phi \wedge \overline{\varrho}$ is true in S and $\varepsilon \in E$. Its application yields the pair (S', E') where $S' = (S \setminus \{\overline{\varrho}\}) \cup \{\varrho\}$ and $E' = (E \setminus \{\overline{\varrho}\}) \cup \{\varrho\}$. ■

In words, a causal relationship is applicable if the associated condition Φ holds, the particular indirect effect ϱ is currently false, and its cause ε is among the current effects. If $\mathcal{R}$ is a set of causal relationships, then by $(S, E) \rightsquigarrow_{\mathcal{R}} (S', E')$ we denote the existence of an element in $\mathcal{R}$ whose application to (S, E) yields (S', E'). Notice that if S is a state and E is consistent, then $(S, E) \rightsquigarrow_{\mathcal{R}} (S', E')$ implies that S' is a state and E' is consistent, too. We adopt a standard notation in writing $(S, E) \overset{*}{\rightsquigarrow}_{\mathcal{R}} (S', E')$ to indicate that there are causal relationships in $\mathcal{R}$ whose successive application to (S, E) yields (S', E').

Example 1 (continued) The following two causal relationships state respectively that the effect *pot* always gives rise to the indirect effect *clog*, and that the effect $\overline{clog}$ (as a result of clearing the tail pipe, say) always gives rise to the indirect effect $\overline{pot}$:[7]

$$\begin{array}{lllll} pot & \texttt{causes} & clog & \texttt{if} & \top \\ \overline{clog} & \texttt{causes} & \overline{pot} & \texttt{if} & \top \end{array} \qquad (4)$$

Recall, now, the state $\{\overline{pot}, \overline{clog}, \overline{runs}, \overline{heavy}\}$ and action *put-p*. Applying the second action law in (2) yields the state $S = \{pot, \overline{clog}, \overline{runs}, \overline{heavy}\}$ along with the effect $E = \{pot\}$. Given the pair (S, E), the first causal relationship in (4) is applicable on account of both $\top \wedge \overline{clog}$ being true in S and $pot \in E$. The application of this relationship yields the pair $((S \setminus \{\overline{clog}\}) \cup \{clog\}, (E \setminus \{\overline{clog}\}) \cup \{clog\})$, i.e.,

$$(\,\{pot, clog, \overline{runs}, \overline{heavy}\}\,,\, \{pot, clog\}\,) \qquad (5)$$

■

Now, suppose given a set of fluent literals S as the result of having computed the direct effects of an action via Definition 2. State S may violate the domain constraints. We then compute additional, indirect effects

[7]See [Thielscher, 1997] on how a suitable set of causal relationships can be automatically extracted from domain constraints given additional knowledge as to which fluents may possibly affect each other.

by (non-deterministically) selecting and (serially) applying causal relationships. If this eventually results in a state satisfying the domain constraints, then this state is considered a *successor* state.

Definition 5 Let $\mathcal{F}$ and $\mathcal{A}$ be sets of fluent and action names, respectively, $\mathcal{L}$ a set of action laws, $\mathcal{D}$ a set of domain constraints, and $\mathcal{R}$ a set of causal relationships. Furthermore, let S be a state satisfying $\mathcal{D}$ and $a \in \mathcal{A}$. A state S' is a *successor state* of S and a iff there exists an applicable (wrt. S) action law $\langle C, a, E\rangle \in \mathcal{L}$ such that

1. $((S \setminus C) \cup E, E) \stackrel{*}{\rightsquigarrow}_{\mathcal{R}} (S', E')$ for some E', and
2. S' satisfies $\mathcal{D}$. ■

Recall, for instance, the state-effect pair in (5). By virtue of satisfying our domain constraint, $pot \supset clog$, its first component constitutes a successor state of $\{\overline{pot}, \overline{clog}, \overline{runs}, \overline{heavy}\}$ and *put-p*. The analogue holds for the Tail Pipe Marauder scenario (Example 2): There are two successor states of $\{\overline{pot}, \overline{clog}, \overline{runs}\}$ and *wait*, viz. $\{\overline{pot}, \overline{clog}, \overline{runs}\}$ and $\{pot, clog, \overline{runs}\}$.

Based on Definition 5, a set of causal laws along with a set of domain constraints and a set of causal relationships determines a *causal model* Σ which maps any pair of an action name and a state to a set of states as follows: $\Sigma(a, S) := \{S' : S' \text{ successor of } S \text{ and } a\}$.

It is important to realize that neither uniqueness nor the existence of a successor state is guaranteed in general; that is, $\Sigma(a, S)$ may contain several elements or may be empty. The former characterizes actions with non-deterministic behavior even though these actions might be deterministic as regards their direct effects. If no successor exists although an applicable action law can be found, then this indicates that the action under consideration has *implicit* preconditions which are not met. While causal relationships account for these qualifications, which derive from domain constraints (see [Thielscher, 1997] for details), notice, however, that implicit preconditions still are strict and as such not part of the qualification problem dealing with the necessity of assuming away abnormal disqualifications.

3 ABNORMAL DISQUALIFICATIONS

We now take the action theory introduced in the preceding section as the basis for our formal account of the qualification problem. The general objective is to appropriately interpret a given formal scenario description and to draw reasonable conclusions about it. Any such description involves general action laws in conjunction with causal relationships, plus specific observations as to both the values of certain fluents and, especially, the non-executability of certain actions in particular situations. The term "reasonable conclusions" appeals to what common sense suggests as to how the given observations are to be interpreted. Formally, a *domain description* (or *domain*, for short) consists of sets $\mathcal{F}$ and $\mathcal{A}$ of fluent and action names; sets $\mathcal{L}$, $\mathcal{D}$, and $\mathcal{R}$ of action laws, domain constraints, and causal relationships, respectively; and a set $\mathcal{O}$ of so-called observations:

Definition 6 Let $\mathcal{F}$ and $\mathcal{A}$ be sets of fluent and action names, respectively. An *observation* is an expression of one of the following forms:

$$F \texttt{ after } [a_1, \ldots, a_n] \quad (6)$$

$$a \texttt{ disqualified after } [a_1, \ldots, a_n] \quad (7)$$

where F is a fluent formula and $a, a_1, \ldots, a_n$ are action names ($n \geq 0$). ■

Intuitively, observation (6) indicates that if the sequence of actions $[a_1, \ldots, a_n]$ were performed in the initial state, then F would hold in the resulting state. Likewise, (7) indicates that after performing the sequence of actions $[a_1, \ldots, a_n]$, action a would be unqualified. For instance, these are possible observations in the context of Example 1:

$$\overline{pot} \wedge \overline{runs} \texttt{ after } [\,]$$
$$start \texttt{ disqualified after } [put\text{-}p]$$

In the remainder of this section, we develop formal notions of interpretations and models for domain descriptions, and we introduce a suitable preference relation among models to allow for assuming away, by default, abnormal disqualifications. This model preference criterion induces a nonmonotonic entailment relation. Together these concepts constitute our proposal how to formalize the qualification problem.

3.1 PERSISTENCE OF ACTION QUALIFICATIONS

The unintended model which occurs in the Put Potato In Tail Pipe scenario when globally minimizing abnormal disqualifications illustrates the necessity of distinguishing disqualifications that admit a *causal* explanation. We have already argued that this can be accomplished by considering the proposition whether an action is or is not abnormally disqualified as potentially being affected by the execution of other actions and otherwise being subject to the general law of persistence. In other words, this proposition is taken as a fluent. According to the general assumption that the world is 'normal' unless there is information to the contrary, this fluent is assumed *initially* false by default. Restricting the assumption of normality to the initial state enables us to consider it normal, as intended, when an action occurs whose effects suggest an action disqualification which, under general circumstances, would be abnormal. Formally, let, for each

action name a, $disq(a)$ be a fluent name. The intended meaning is that if $\overline{disq(a)}$ holds in some state, then action a is not disqualified for some abnormal reason—which shall imply that a be qualified if and only if all strict preconditions are satisfied.[8]

Abnormal disqualifications indicate abnormal circumstances. These may be described by fluents which, too, are to be assumed false by default. Example fluents of this kind might be *clog* and *pot*, as one normally assumes that the tail pipe is not clogged, let alone the possibility of its housing a potato. Fluents denoting abnormal circumstances can be combined in domain constraints to describe the conditions for an action being abnormally disqualified. In particular, it is often desirable to equate a fluent $disq(a)$ with a disjunction consisting of all (to the best of the agent's knowledge) the causes for an abnormal disqualification of a. This does not only allow to derive an action disqualification from the occurrence of one of its causes, it also supports the proliferation of explanations for abnormal disqualifications that have been observed (see Section 3.2.2, below).

To make all this precise, let $\mathcal{F}$ and $\mathcal{A}$ be the sets of fluent and action names, respectively, of a domain description. From now on we always assume determined a certain subset $\mathcal{F}_{ab} \subseteq \mathcal{F}$ of fluents that will be considered initially false by default. It is moreover assumed that $disq(a) \in \mathcal{F}_{ab}$ for each action name $a \in \mathcal{A}$. A typical domain constraint, then, is of the form

$$disq(a) \equiv \bigvee_{i \in I_a} f_i \tag{8}$$

for some index set I_a such that each $f_i \in \mathcal{F}_{ab}$. That is, each of the 'abnormality' fluents f_i is a potential cause of an abnormal disqualification of action a.[9] These domain constraints may give rise to indirect effects, namely, a change of the truth value of an element in the disjunction might also affect the truth value of $disq(a)$.

Example 1 (continued) Let the set $\mathcal{F}_{ab}$ consist of the fluents *pot*, *clog*, *heavy*, along with $disq(start)$ and $disq(put\text{-}p)$. Suppose further that the set of domain constraints includes

$$\begin{aligned} disq(start) &\equiv clog \\ disq(put\text{-}p) &\equiv heavy \end{aligned} \tag{9}$$

aside from $pot \supset clog$. The additional domain constraints are accompanied by these causal relationships:

$$\begin{aligned} \overline{clog} &\ \texttt{causes}\ \overline{disq(start)} \ \texttt{if}\ \top \\ clog &\ \texttt{causes}\ disq(start) \ \texttt{if}\ \top \\ \overline{heavy} &\ \texttt{causes}\ \overline{disq(put\text{-}p)} \ \texttt{if}\ \top \\ heavy &\ \texttt{causes}\ disq(put\text{-}p) \ \texttt{if}\ \top \end{aligned} \tag{10}$$

in conjunction with the ones shown in (4). Suppose, now, action *put-p* is performed in the state $S = \{\overline{pot}, \overline{clog}, \overline{runs}, \overline{heavy}, \overline{disq(start)}, \overline{disq(put\text{-}p)}\}$. The application of the corresponding action law in (2) yields the state-effect pair

$$(\{pot, \overline{clog}, \overline{runs}, \overline{heavy}, \overline{disq(start)}, \overline{disq(put\text{-}p)}\}, \{pot\})$$

The first component does not satisfy $pot \supset clog$, but we can apply the first causal relationship in (4), viz. *pot* `causes` *clog* `if` $\top$, yielding

$$(\{pot, clog, \overline{runs}, \overline{heavy}, \overline{disq(start)}, \overline{disq(put\text{-}p)}\}, \{pot, clog\})$$

While now the aforementioned domain constraint is satisfied, the first fluent formula in (9) is no longer so. Yet we can further apply the appropriate causal relationship in (10), viz. *clog* `causes` $disq(start)$ `if` $\top$, which results in

$$(\{pot, clog, \overline{runs}, \overline{heavy}, disq(start), \overline{disq(put\text{-}p)}\}, \{pot, clog, disq(start)\}) \tag{11}$$

This pair's first component satisfies all domain constraints and, thus, constitutes a successor state. Notice that action *start* is declared abnormally disqualified in the resulting state. This disqualification occurs as an indirect effect of having performed *put-p*. On the other hand, executing this action did not affect the fluent $disq(put\text{-}p)$, which thus remains false according to the law of persistence. ■

3.2 ASSUMING QUALIFICATION BY DEFAULT

The intention of distinguishing a set of 'abnormality' fluents $\mathcal{F}_{ab}$ is to prefer among all suitable interpretations of domain descriptions those in which they are initially false. This would enable us to assume away abnormal circumstances whenever that is reasonable. Prior to discussing preference, however, we need to formalize the general notions of interpretation and model. Clearly, they both ought to respect the causal model Σ underlying the domain in question. Each interpretation (and model) contains a partial function *Res* which maps finite action sequences to states with the intended meaning that $Res([a_1, \ldots, a_n])$ would be the result of executing the action sequence $[a_1, \ldots, a_n]$ in the initial state (which itself is determined by $Res([\,])$).

[8] For the moment we neglect the possibility of miraculous disqualifications, which will be discussed later, in Section 3.3.

[9] Instead of explicitly providing the "only-if" part in (8), i.e., $disq(a) \supset \bigvee_{i \in I_a} f_i$, this could be implicitly obtained through circumscribing [McCarthy, 1980] the predicate *disq* in a given set of domain constraints; c.f. [Lifschitz, 1987], where this idea is applied to strict preconditions of actions.

Definition 7 Let Σ be the causal model determined by a domain description with domain constraints $\mathcal{D}$. A pair (Res, Σ) is an *interpretation* for this domain iff Res is a partial mapping from finite sequences of action names to states such that the following holds:

1. $Res([\,])$ is defined and satisfies $\mathcal{D}$.
2. For any finite sequence $[a_1, \ldots, a_{n-1}, a_n]$ of action names ($n > 0$), $Res([a_1, \ldots, a_{n-1}, a_n])$ is defined iff
   (a) $Res([a_1, \ldots, a_{n-1}])$ is defined;
   (b) $\overline{disq(a_n)} \in Res([a_1, \ldots, a_{n-1}])$; and
   (c) $\Sigma(a_n, Res([a_1, \ldots, a_{n-1}])) \neq \{\}$

   If it is defined, then $Res([a_1, \ldots, a_{n-1}, a_n])$ is a successor of $Res([a_1, \ldots, a_{n-1}]))$ and a_n. ■

If $Res([a_1, \ldots, a_n])$ is defined, we also say that the action sequence $[a_1, \ldots, a_n]$ is *qualified*. Then Definition 7 states that $[a_1, \ldots, a_{n-1}, a_n]$ is qualified if so is $[a_1, \ldots, a_{n-1}]$, if the state $Res([a_1, \ldots, a_{n-1}])$ does not imply an abnormal disqualification of a_n — which is indicated by fluent $disq(a_n)$ being false in this state—, and if all strict preconditions of a_n are met—which implies the existence of a successor state of a_n and $Res([a_1, \ldots, a_{n-1}])$. Notice that all defined function values of Res necessarily satisfy the underlying domain constraints if $Res([\,])$ does.

Based on the given a set of observations, an interpretation for a domain is considered a model iff all the observations hold in that interpretation.

Definition 8 Let Σ be the causal model of a domain description with observations $\mathcal{O}$. An interpretation (Res, Σ) is a *model of* $\mathcal{O}$ iff each observation in $\mathcal{O}$ holds in (Res, Σ), where

1. F `after` $[a_1, \ldots, a_n]$ is said to *hold* in (Res, Σ) iff $Res([a_1, \ldots, a_n])$ is defined and F is true in $Res([a_1, \ldots, a_n])$;
2. a `disqualified after` $[a_1, \ldots, a_n]$ is said to *hold* in (Res, Σ) iff $Res([a_1, \ldots, a_n])$ is defined but $Res([a_1, \ldots, a_n, a])$ is not. ■

Example 1 (continued) Let Σ be the causal model determined by the action laws (2), the domain constraints (1) and (9), and the causal relationships (4) and (10). Suppose given the observation

$$\overline{runs} \texttt{ after } [\,] \tag{12}$$

and consider, say, these two initial states:

$$\begin{aligned} Res_1([\,]) &= \{\overline{pot}, \overline{clog}, \overline{runs}, \overline{heavy}, \\ &\quad \overline{disq(start)}, \overline{disq(put\text{-}p)}\} \\ Res_2([\,]) &= \{\overline{pot}, \overline{clog}, \overline{runs}, heavy, \\ &\quad \overline{disq(start)}, disq(put\text{-}p)\} \end{aligned} \tag{13}$$

The corresponding interpretations[10] (Res_1, Σ) and (Res_2, Σ) satisfy (12), hence are models. Notice, however, that no 'abnormality' fluent is true in $Res_1([\,])$, as opposed to $Res_2([\,])$. Since $disq(start)$ holds in $\Sigma(put\text{-}p, Res_1([\,]))$ (c.f. (11)), the model (Res_1, Σ) entails that the engine cannot be ignited after putting a potato into the tail pipe. In contrast, the model (Res_2, Σ) is the formal counterpart of the counterintuitive conclusion where the action $put\text{-}p$ is assumed to be abnormally disqualified in the first place. ■

While an interpretation must satisfy the given observations in order to constitute a model, this criterion alone does not suffice to assume away abnormal disqualifications. Obviously, the addition of observations can only decrease the set of models, never produce new ones. Consequently, if one defines an entailment relation stating that an observation is entailed by a set of observations if the former holds in all models of the latter, then this relation is monotone. Under the name *restricted monotonicity*, in [Lifschitz, 1993] this property is claimed generally desirable in theories of actions. Yet this is no longer appropriate when being confronted with the qualification problem because additional observations, such as detecting a potato in the tail pipe, may force us to withdraw previous (default) conclusions, like the conclusion that we are able to start the engine. We achieve this formally by a preference relation among the set of models, with the intention to select those which initially minimize truth of fluents in $\mathcal{F}_{ab}$ to the largest possible extent. When talking about entailment, attention is then restricted to models which are preferred in this sense. The following definition constitutes the core of our formal characterization of the qualification problem:

Definition 9 Let $\mathcal{F} \supseteq \mathcal{F}_{ab}$ be the underlying set of fluent names and $\mathcal{O}$ the set of observations of a domain description with causal model Σ. An interpretation $M' = (Res', \Sigma)$ is *less abnormal* than an interpretation $M = (Res, \Sigma)$, written $M' \prec M$, iff $Res'([\,]) \cap \mathcal{F}_{ab} \subsetneqq Res([\,]) \cap \mathcal{F}_{ab}$.

A model M of $\mathcal{O}$ is *preferred* iff there is no model M' of $\mathcal{O}$ such that $M' \prec M$. An observation o is *entailed*, written $\mathcal{O} \mid\!\sim_\Sigma o$, iff o holds in each preferred model of $\mathcal{O}$. ■

In words, the less fluents in $\mathcal{F}_{ab}$ occur affirmatively in the initial state in a model the better. Obviously, the induced entailment relation, $\mid\!\sim_\Sigma$, is nonmonotonic as the addition of observations may change the set of preferred models entirely. In the sequel, we illustrate how this formal account of the qualification problem satisfies all the requirements which we demanded in the introduction.

[10] Notice that if all actions in a domain are deterministic (that is, each $\Sigma(a, S)$ is singleton or empty), then an interpretation (Res, Σ) is uniquely characterized by its initial state, $Res([\,])$.

3.2.1 How To Assume Away Disqualifications

The fundamental issue with the qualification problem is to assume away abnormal disqualifications by default. This, however, should concern only those disqualifications which do not admit a causal explanation. Our key example, in particular, is now treated in the expected way. Namely, any potential abnormal disqualification preventing us from putting a potato into the tail pipe is assumed away, for there is no evidence to the contrary. Likewise, any abnormal disqualification preventing us from starting the engine is assumed away as regards the initial state, whereas an abnormal disqualification of this very action after the insertion of a potato follows from the causal model without the necessity of granting abnormal circumstances.

Example 1 (continued) Recall from (13) the two models $M_1 = (Res_1, \Sigma)$ and $M_2 = (Res_2, \Sigma)$ of (12). Clearly, we have $M_1 \prec M_2$ due to $Res_1([\,]) \cap \mathcal{F}_{ab} = \{\}$ and $Res_2([\,]) \cap \mathcal{F}_{ab} = \{\mathit{heavy}, \mathit{disq}(\mathit{put\text{-}p})\}$. Since each 'abnormality' fluent is false in the initial state in M_1, the latter obviously constitutes the unique preferred model. Whatever holds in M_1 is thus entailed by the domain. In particular, we have seen in (11) that $\mathit{disq}(\mathit{start}) \in Res_1([\mathit{put\text{-}p}])$. This implies that $[\mathit{put\text{-}p}, \mathit{start}]$ is not qualified in M_1, which in turn sanctions the entailment of

$$\mathit{start} \ \texttt{disqualified after} \ [\mathit{put\text{-}p}]$$

This constitutes the intended solution: The first action, *put-p*, is qualified by default and, as a consequence, action *start* is unqualified afterwards. ■

3.2.2 How To Explain Disqualifications

Aside from assuming away abnormal disqualifications of actions by default, one naturally seeks conceivable explanations in case a disqualifications has been—unexpectedly—observed without an apparent cause. Each preferred model that contains an abnormal disqualification also includes, provided the underlying domain constraints support this, a particular explanation. For otherwise the domain constraints would be violated in the state in which the disqualification occurs, as the following example illustrates.

Example 3 We extend the set of fluent names of Example 1 by *no-gas*, *low-batt*, and *engine-problem*, each of which shall belong to the subset $\mathcal{F}_{ab}$. These fluent names are combined in this domain constraint:

$$\mathit{disq}(\mathit{start}) \equiv \mathit{clog} \vee \mathit{no\text{-}gas} \vee \mathit{low\text{-}batt} \vee \mathit{engine\text{-}problem}$$

which shall replace the first formula in (9). Now suppose we are in a state where the engine is not running and where we also know that the tail pipe is not clogged nor is the tank empty, but nonetheless we encounter difficulties with starting the engine. The corresponding observations, i.e.,

$$\overline{\mathit{runs}} \ \texttt{after} \ [\,]$$
$$\overline{\mathit{clog}} \wedge \overline{\mathit{no\text{-}gas}} \ \texttt{after} \ [\,]$$
$$\mathit{start} \ \texttt{disqualified after} \ [\,]$$

admit two preferred models: Each model (Res, Σ) must satisfy $\mathit{disq}(\mathit{start}) \in Res([\,])$ since $[\mathit{start}]$ is unqualified, according to the third observation, although the only strict precondition of *start*, viz. $\overline{\mathit{runs}}$, is initially true according to the first observation. Given $\mathit{disq}(\mathit{start}) \in Res([\,])$, the above domain constraint requires an additional 'abnormality' fluent be initially true in any model. The second observation excludes both *clog* and *no-gas*. Hence, a preferred model satisfies either $\mathit{low\text{-}batt} \in Res([\,])$ or else $\mathit{engine\text{-}problem} \in Res([\,])$. This in turn sanctions the entailment of the observation

$$\mathit{low\text{-}batt} \vee \mathit{engine\text{-}problem} \ \texttt{after} \ [\,] \qquad (14)$$

That is, problems with the battery or problems with the engine explain the observed abnormal disqualification of *start*. ■

3.2.3 How To Deal With Non-Determinism

The failure of the *chronological ignorance* approach to the qualification problem [Shoham, 1987; Shoham, 1988] in case of non-deterministic actions demonstrates a crucial difficulty with combining both abnormal disqualifications and non-determinism. The problem occurs whenever non-deterministic information provides sufficient evidence for an abnormal disqualification without, by virtue of being non-deterministic, necessitating it. Any formalism by which abnormal circumstances are negated whenever they do not provably hold, ignores uncertain evidence and, in so doing, supports unsound conclusions. As the Tail Pipe Marauder example will illustrate, our formal characterization of the qualification problem does not interfere with non-deterministic information and treats the latter in the appropriate, namely, the cautious way.

Example 2 (continued) Suppose given the observation

$$\overline{\mathit{runs}} \ \texttt{after} \ [\,]$$

Since it is consistent with the observation to consider initially false all members of $\mathcal{F}_{ab}$, any preferred model (Res, Σ) must satisfy

$$Res([\,]) = \{\overline{\mathit{pot}}, \overline{\mathit{clog}}, \overline{\mathit{runs}}, \overline{\mathit{disq}(\mathit{wait})}, \overline{\mathit{disq}(\mathit{start})}\}$$

The action *wait* being non-deterministic (c.f. (3)), we know that either $Res([\mathit{wait}]) = Res([\,])$ or else $Res([\mathit{wait}]) = \{\mathit{pot}, \mathit{clog}, \overline{\mathit{runs}}, \overline{\mathit{disq}(\mathit{wait})}, \mathit{disq}(\mathit{start})\}$ holds in preferred models. Therefore, nothing definite follows about the status of the tail pipe, hence of the qualification of *start*, after performing $[\mathit{wait}]$. Consequently, the observation *runs* `after` $[\mathit{wait}, \mathit{start}]$, say, is not entailed, as intended. ■

3.3 MIRACULOUS DISQUALIFICATIONS

Thus far our theory supports generating explanations for surprising disqualifications by selecting among the conceivable reasons for this abnormality. Yet whenever the domain description renders invalid each of these explanations, then that goes beyond the capacity of the theory. Suppose given, as an example, the two observations

$$\begin{array}{l} \mathit{start} \ \texttt{disqualified after} \ [\,] \\ \mathit{runs} \ \texttt{after} \ [\mathit{wait}, \mathit{start}] \end{array} \tag{15}$$

where *wait* is assumed to have no effects at all on the underlying fluents. No however (*a priori*) 'unlikely' model exists which simultaneously satisfies both of the observations. The reason is that any abnormality explaining the first disqualification necessarily transfers to the state after waiting, which contradicts the following success of performing *start*. Nonetheless, such situations, where the available explanations are insufficient to account for surprising disqualifications, are well conceivable and just prove our lack of omniscience.

We therefore need to extend our formalism to allow for observed yet inexplicable, in the above sense, action disqualifications. To this end, the formal notions of interpretation and model are enhanced by a component accommodating these so-called *miraculous* disqualifications. As we have seen, a miraculous disqualification may appear or disappear even though the truth values of the fluents suggest identical states. This is why any such disqualification is to be associated with the sequence of actions after whose execution it occurs, rather than with the respective state. Formally, the new component, denoted by Υ, consists of non-empty action sequences indicating the following: Whenever $[a_1, \ldots, a_{n-1}, a_n] \in \Upsilon$ $(n > 0)$, then action a_n is disqualified in the state resulting from performing $[a_1, \ldots, a_n]$ even if all strict preconditions of a_n and also $\overline{\mathit{disq}(a_n)}$ hold in that state. The following extends Definition 7 accordingly.

Definition 10 Let Σ be the causal model determined by a domain description with domain constraints $\mathcal{D}$. A triple $(\mathit{Res}, \Sigma, \Upsilon)$ is an *interpretation* for this domain iff Υ is a set of non-empty, finite sequences of action names and *Res* is a partial mapping from finite sequences of action names to states such that the following holds:

1. $\mathit{Res}([\,])$ is defined and satisfies $\mathcal{D}$.
2. For any finite sequence $[a_1, \ldots, a_{n-1}, a_n]$ of action names $(n > 0)$, $\mathit{Res}([a_1, \ldots, a_{n-1}, a_n])$ is defined iff
   (a) $\mathit{Res}([a_1, \ldots, a_{n-1}])$ is defined;
   (b) $\overline{\mathit{disq}(a_n)} \in \mathit{Res}([a_1, \ldots, a_{n-1}])$;
   (c) $\Sigma(a_n, \mathit{Res}([a_1, \ldots, a_{n-1}])) \neq \{\}$; and
   (d) $[a_1, \ldots, a_{n-1}, a_n] \notin \Upsilon$.

   If it is defined, then $\mathit{Res}([a_1, \ldots, a_{n-1}, a_n])$ is a successor of $\mathit{Res}([a_1, \ldots, a_{n-1}]))$ and a_n. ■

The additional clause, 2(d), states that a sequence of actions can only be qualified if it is not miraculously disqualified. As before, a model of a set of observations is an interpretation in which all the observations hold (c.f. Definition 8).

Example 4 The domain discussed in Example 1 is extended by the action name *wait* in conjunction with the action law $\langle\{\}, \mathit{wait}, \{\}\rangle$. Furthermore, suppose given the aforementioned observations (15). While no model $(\mathit{Res}, \Sigma, \Upsilon)$ with $\Upsilon = \{\}$ exists for this domain, as argued above, both these observations hold in the interpretation $(\mathit{Res}, \Sigma, \Upsilon)$ where $\mathit{Res}([\,])$ is

$$\{\overline{\mathit{pot}}, \overline{\mathit{clog}}, \overline{\mathit{runs}}, \overline{\mathit{heavy}}, \overline{\mathit{disq}(\mathit{start})}, \overline{\mathit{disq}(\mathit{put\text{-}p})}\} \tag{16}$$

and $\Upsilon = \{[\mathit{start}]\}$. This interpretation thus constitutes a model. ■

Clearly, miraculous disqualifications, too, are to be minimized to the largest possible extent. Moreover, miraculous disqualifications are meant as means to account for abnormal disqualifications which do not admit an explanation even by granting abnormal circumstances. As such, miraculous disqualifications need to be minimized with higher priority. As opposed to explicable disqualifications, miraculous ones can well be minimized globally, that is, without worrying about causality—would they admit a causal explanation they would not be miraculous. We thus arrive at the following extension of our preference criterion:

Definition 11 Let $\mathcal{F} \supseteq \mathcal{F}_{ab}$ be the underlying set of fluent names and $\mathcal{O}$ the set of observations of a domain description with causal model Σ. An interpretation $M' = (\mathit{Res}', \Sigma, \Upsilon')$ is *less abnormal* than an interpretation $M = (\mathit{Res}, \Sigma, \Upsilon)$, written $M' \prec M$, iff

1. either $\Upsilon' \subsetneqq \Upsilon$,
2. or $\Upsilon' = \Upsilon$ and $\mathit{Res}'([\,]) \cap \mathcal{F}_{ab} \subsetneqq \mathit{Res}([\,]) \cap \mathcal{F}_{ab}$.

The notions of preferred model and entailment in Definition 9 modify accordingly. ■

Example 4 (continued) We have seen that the domain considered above does not admit a model without miraculous disqualifications. It follows that the above model $M = (\mathit{Res}, \Sigma, \Upsilon)$—where $\mathit{Res}([\,])$ is as in (16) and $\Upsilon = \{[\mathit{start}]\}$—is preferred, for it declares a single action sequence miraculously disqualified and negates each 'abnormality' fluent in the initial state. As a matter of fact, M is the only preferred model since any model $(\mathit{Res}', \Sigma, \Upsilon')$ must satisfy $[\mathit{start}] \in \Upsilon'$ and also $\overline{\mathit{runs}} \in \mathit{Res}'([\,])$ (the latter is due to $[\mathit{wait}, \mathit{start}]$ being qualified according to (15)). ■

4 FLUENT CALCULUS AND THE QUALIFICATION PROBLEM

Finally, we briefly sketch a suitable action calculus which is capable of handling abnormal action disqualifications. Our encoding employs the representation technique underlying the *fluent calculus* [Hölldobler and Schneeberger, 1990; Thielscher, 1997]. As opposed to the situation calculus [McCarthy and Hayes, 1969; Reiter, 1991], the fluent calculus employs structured state terms, each of which consists in a collection of all fluent literals that are true in the state being represented. To this end, fluent literals are reified, i.e., formally represented as terms. An example state term is $\mathit{in\text{-}pipe}(\mathit{potato}) \circ \overline{\mathit{heavy}(\mathit{potato})} \circ \mathit{clog}$ [11] where the bar denoting negative fluent expressions is formally a unary function and where $\circ$ denotes a special binary function symbol which obeys the laws of associativity and commutativity. It has first been argued in [Hölldobler and Schneeberger, 1990] that this representation technique avoids extra axioms (e.g., frame axioms [McCarthy and Hayes, 1969]) to encode the general law of persistence: The effects of actions are modeled by manipulating state terms through removal and addition of sub-terms. Then all sub-terms which are not affected by these operations remain in the state term, hence continue to be true.

Our solution to the qualification problem in the fluent calculus builds on the integration of causal relationships into this calculus [Thielscher, 1997]. While the fluent calculus provides monotonic solutions to both the frame problem as well as the ramification problem, the qualification problem, as we have seen, necessitates some kind of nonmonotonicity. In particular, we employ for each 'abnormality' fluent name $f_{ab} \in \mathcal{F}_{ab}$ the *default rule* [Reiter, 1980]

$$\frac{: \forall s\,[\,\mathit{Initial}(s) \supset \neg \mathit{Holds}(f_{ab}(x_1, \ldots, x_n), s)\,]}{\forall s\,[\,\mathit{Initial}(s) \supset \neg \mathit{Holds}(f_{ab}(x_1, \ldots, x_n), s)\,]}$$

This rule should be read as: Provided it is consistent, conclude that if s represents the initial state then an instance $f_{ab}(t_1, \ldots, t_n)$ is false in s. In addition, miraculous disqualifications are assumed away, whenever possible, by applying defaults of the form

$$\frac{: \neg \mathit{Miracle}(a^*)}{\neg \mathit{Miracle}(a^*)} \quad (\ a^*\ \text{action sequence})$$

Since miraculous disqualifications are to be minimized with higher priority, we employ the concepts of *Prioritized Default Logic* [Brewka, 1994]. The report [Thielscher, 1996] contains full details as well as a formal proof of the adequacy of this extension with regard to the theory developed in Section 3.

[11] As opposed to the formal language used in the preceding sections, our action calculus supports non-propositional fluents, such as *in-pipe*, whose arguments are chosen from a set of *entities*, such as *potato*.

5 DISCUSSION

We have proposed a formal characterization of the qualification problem from the perspective that requiring global minimization of abnormal disqualifications is obviously inadequate. Our theory may be summarized as follows. Any domain description is supposed to contain a distinguished set of fluents $\mathcal{F}_{ab}$, each of which describes abnormal circumstances and thus is to be assumed false by default. This assumption, however, needs to be restricted to the initial state, so that these fluents are subject to the general law of persistence but are also potentially (directly or indirectly) affected by the execution of actions. Among these 'abnormality' fluents are propositions, denoted $\mathit{disq}(a)$, which state that an action a is abnormally disqualified. Domain constraints relating these fluents with possible causes of an abnormal disqualification support the proliferation of explanations in case an abnormal disqualification—surprisingly—occurs. In addition, miraculous disqualifications accommodate situations in which a suitable explanation cannot be provided. The default assumption of 'normality' is formally represented by a model preference criterion (Definition 11), which induces a nonmonotonic entailment relation among observations.

Using a suitably simple action language, the focus in this paper has been on the qualification problem. The underlying principles of our theory, however, are sufficiently fundamental and general to not depend on this specific language. Thus these principles could equally well be employed in other, more elaborated formal theories of actions like, e.g., [Gelfond and Lifschitz, 1993; Sandewall, 1994; Thielscher, 1995], in view of the qualification problem. Likewise, existing action calculi may be enhanced on this basis in order that they become capable of dealing with abnormal action disqualifications. As an example, we have sketched a way to embed the fluent calculus in an appropriate nonmonotonic theory. The adequacy of the resulting framework has been established by relating it to our formal characterization of the qualification problem. This adds another item to the list of ontological aspects which the fluent calculus is capable of dealing with, such as non-deterministic and concurrent actions [Bornscheuer and Thielscher, 1997], indirect effects of actions [Thielscher, 1997], and continuous change [Herrmann and Thielscher, 1996].

Besides the proposal pursued in this paper, the only existing alternative to global minimization of abnormalities as a solution to the qualification problem is the concept of *chronological ignorance* [Shoham, 1987; Shoham, 1988]. Roughly speaking, the crucial idea there is to assume away, by default, abnormal circumstances which do not provably hold, and simultaneously to prefer minimization of abnormalities at earlier timepoints. This approach treats our introductory key example correctly. The interesting, albeit informal,

reason for coming to the desired conclusion in this and similar cases is a certain respect of causality hidden in this method: By minimizing chronologically, one tends to minimize causes rather than effects—which is the right thing to do—simply because in general causes precede effects. On the other hand, it has already been shown elsewhere (e.g., [Kautz, 1986; Sandewall, 1993; Stein and Morgenstern, 1994]) that the applicability of chronological minimization is intrinsically restricted to domains which do not include non-deterministic information. The Tail Pipe Marauder scenario of Example 2 constitutes a simple domain which does not fall into that category. Given that non-deterministically there might or might not be a potato in the tail pipe, chronological ignorance sanctions the prediction that nonetheless starting the engine will be successful. For it cannot be *proved* that this action has an abnormal disqualification—which thus is assumed away. While the qualification problem means to assume away abnormal circumstances whenever they do not provably hold, the Tail Pipe Marauder domain illustrates that this approach is in general too optimistic if the execution of a non-deterministic action renders quite possible such circumstances. In contrast, our characterization of the qualification problem accounts for this as the minimization procedure applied to abnormal or miraculous disqualifications does not interfere with the results of non-deterministic actions.

Our approach to the qualification problem shares with *Motivated Action Theory* [Stein and Morgenstern, 1994] the insight that an appropriate notion of causality is necessary when assuming away abnormalities. In the latter framework, occurrences of actions and events are assumed away by default while considering the possibility that they are caused (or, in other words, *motivated*, hence the name). This minimizing unmotivated events and our minimizing non-caused abnormal disqualifications are somehow complementary while based on similar principles. Of course, the formal realizations are quite different. An unsatisfactory property of Motivated Action Theory is that the preference criterion, that is, *motivation*, depends on the syntactical structure of the formulas representing causal knowledge. As a consequence, logical equivalent formalizations may induce different preference criteria, of which only one is the desired. Moreover, the formal concept of motivation becomes rather complicated in case of disjunctive (i.e., non-deterministic) information, which entails difficulties with assessing its range of applicability.

Throughout the paper, we have taken action disqualifications as rendering physically impossible the execution of the respective action. A desirable refinement is to consider actions be disqualified *as to producing a certain effect* (c.f. [Gelfond *et al.*, 1991], e.g.). This is accomplished with a simple, straightforward extension of our theory. In addition to the fluents $\mathit{disq}(a)$, we introduce fluents of the form $\mathit{disq}(a, \ell)$, whose intended reading is "action a fails to produce effect ℓ." These fluents, too, belong to the set $\mathcal{F}_{ab}$ and may be related to other 'abnormality' fluents by means of domain constraints, like in

$$\mathit{disq}(\mathit{shoot}, \overline{\mathit{alive}}) \equiv \mathit{bad\text{-}sight} \vee \mathit{bad\text{-}shooter} \vee \mathit{bad\text{-}gun}$$

Suppose, then, $\langle C, a, E\rangle$ is the action law to be applied to some state S. The effect which a actually manages to produce if performed in S is formally given by $E' := E \setminus \{\ell : \mathit{disq}(a, \ell) \in S\}$. Let $C' := C \setminus \{\ell, \overline{\ell} : \ell \in E \setminus E'\}$, which guarantees that $|C'| = |E'|$, then $(S \setminus C') \cup E'$ is taken as the intermediate state which is subject to the following ramification process. The notion of a successor state modifies accordingly while all further concepts, viz. interpretations, models, and the preference criterion, remain unaltered.

Finally, it needs to be mentioned that we gave emphasis only to the representational aspect of the qualification problem, as opposed to the computational aspect. That the latter is of equal importance has been pointed out, e.g., in [Elkan, 1995]. Our analysis has revealed some hitherto unnoticed problems with the representational aspect and, to state the obvious, the computational aspect cannot be pursued without an appropriate representation of the problem. Named the computational part of the qualification problem, the challenge is to find a computational model that enables the reasoning agent to assume that an action be qualified without even *thinking* of all possible disqualifying causes—unless some piece of knowledge hints at their presence. In principle, the special fluents $\mathit{disq}(a)$ employed in our theory serve this purpose: By assuming $\overline{\mathit{disq}(a)}$, one jumps to the conclusion that a be qualified provided all strict preconditions are met. Still, on the other hand, in order that this assumption be justified, its consistency as regards the underlying domain constraints must be guaranteed. In a standard reasoning system, this in turn involves consideration (and exclusion) of all the potential disqualifying, abnormal circumstances. A solution to the computational part of the qualification problem thus requires a different computational model, presumably based on some parallel architecture, by which all related domain constraints are ignored unless they are explicitly 'activated' by some piece of information. Although this aspect was not among the topics of this paper, the foundations have been laid.

Acknowledgments

The author is grateful to Wolfgang Bibel, Christoph Herrmann, and two anonymous referees for helpful comments and suggestions, and to Erik Sandewall for enlightening remarks on the subject of the paper.

References

[Bornscheuer and Thielscher, 1997]
S. Bornscheuer and M. Thielscher. Explicit and implicit indeterminism: Reasoning about uncertain and contradictory specifications of dynamic systems. *J. of Logic Programming*, 1997.

[Brewka, 1994] G. Brewka. Adding priorities and specificity to default logic. In C. MacNish, D. Pearce, and L. Pereira, eds., *Proc. of the European Workshop on Logics in AI (JELIA)*, vol. 838 of *LNAI*, p. 50–65. Springer 1994.

[Elkan, 1995] C. Elkan. On solving the qualification problem. In C. Boutilier and M. Goldszmidt, eds., *Extending Theories of Actions: Formal Theory and Practical Applications*, vol. SS–95–07 of *AAAI Spring Symposia*, Stanford University, 1995.

[Fikes and Nilsson, 1971] R. Fikes and N. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:189–208, 1971.

[Gelfond and Lifschitz, 1993] M. Gelfond and V. Lifschitz. Representing action and change by logic programs. *J. of Logic Programming*, 17:301–321, 1993.

[Gelfond *et al.*, 1991] M. Gelfond, V. Lifschitz, and A. Rabinov. What are the limitations of the situation calculus? In S. Boyer, ed., *Automated Reasoning, Essays in Honor of Woody Bledsoe*, p. 167–181. Kluwer Academic, 1991.

[Ginsberg and Smith, 1988a] M. Ginsberg and D. Smith. Reasoning about action I: A possible worlds approach. *Artificial Intelligence*, 35:165–195, 1988.

[Ginsberg and Smith, 1988b] M. Ginsberg and D. Smith. Reasoning about action II: The qualification problem. *Artificial Intelligence*, 35:311–342, 1988.

[Hanks and McDermott, 1987] S. Hanks and D. McDermott. Nonmonotonic logic and temporal projection. *Artificial Intelligence*, 33(3):379–412, 1987.

[Herrmann and Thielscher, 1996] C. Herrmann and M. Thielscher. Reasoning about continuous processes. In B. Clancey and D. Weld, eds., *Proc. of the AAAI*, p. 639–644, Portland, 1996.

[Hölldobler and Schneeberger, 1990]
S. Hölldobler and J. Schneeberger. A new deductive approach to planning. *New Generation Computing*, 8:225–244, 1990.

[Kautz, 1986] H. Kautz. The logic of persistence. In *Proc. of the AAAI*, p. 401–405, Philadelphia, 1986.

[Lifschitz, 1987] V. Lifschitz. Formal theories of action (preliminary report). In J. McDermott, ed., *Proc. of the IJCAI*, p. 966–972, Milan, 1987.

[Lifschitz, 1993] V. Lifschitz. Restricted monotonicity. In *Proc. of the AAAI*, p. 432–437, Washington, 1993.

[Lin and Reiter, 1994] F. Lin and R. Reiter. State constraints revisited. *J. of Logic and Computation*, 4(5):655–678, 1994.

[McCarthy and Hayes, 1969] J. McCarthy and P. Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502, 1969.

[McCarthy, 1977] J. McCarthy. Epistemological problems of artificial intelligence. In *Proc. of the IJCAI*, p. 1038–1044, Cambridge, 1977.

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

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

[Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Reiter, 1991] R. Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz, ed., *Artificial Intelligence and Mathematical Theory of Computation*, p. 359–380. Academic Press, 1991.

[Sandewall, 1993] E. Sandewall. Systematic assessment of temporal reasoning methods for use in autonomous systems. In B. Fronhöfer, ed., *Workshop on Reasoning about Action & Change at IJCAI*, p. 21–36, Chambéry, 1993.

[Sandewall, 1994] E. Sandewall. *Features and Fluents. The Representation of Knowledge about Dynamical Systems.* Oxford University Press, 1994.

[Shoham, 1987] Y. Shoham. *Reasoning about Change.* MIT Press, 1987.

[Shoham, 1988] Y. Shoham. Chronological ignorance: Experiments in nonmonotonic temporal reasoning. *Artificial Intelligence*, 36:279–331, 1988.

[Stein and Morgenstern, 1994] L. Stein and L. Morgenstern. Motivated action theory: A formal theory of causal reasoning. *Artificial Intelligence*, 71:1–42, 1994.

[Thielscher, 1995] M. Thielscher. The logic of dynamic systems. In C. Mellish, ed., *Proc. of the IJCAI*, p. 1956–1962, Montreal, 1995.

[Thielscher, 1996] M. Thielscher. Qualification and Causality. Technical Report TR-96-026, ICSI, Berkeley, July 1996. (Available at `http://kirmes.inferenzsysteme.informatik.th-darmstadt.de/~mit`; click the item `Technical Reports`).

[Thielscher, 1997] M. Thielscher. Ramification and causality. *Artificial Intelligence*, 1997. (To appear. A preliminary version is available as Technical Report TR-96-003, ICSI, Berkeley, Jan. 1996).