Ann Math Artif Intell (2006) 46: 375–413
DOI 10.1007/s10472-006-9032-3

A preferential semantics for causal reasoning about action

Mikhail Prokopenko

Received: 9 January 2006 / Published online: 10 October 2006

Abstract One of the principal concerns in the research area of Reasoning about Action is determining the ramifications of actions in changing environments. A particular tendency emerging in recent literature endorses the explicit incorporation of causal knowledge in logic-based action theories. It is argued that causal extensions not only enhance the expressive power of theories of action, but may also provide more concise and intuitive representations. This paper investigates semantics for causal reasoning about action and change. It does so by exploring the role of several fundamental underlying principles, such as the *Principle of Minimal Change* and the *Principle of Causal Change*. This work culminates in a general unifying semantics for a class of action theories represented by a number of recent and influential approaches – in particular, the causal relationship approach of Thielscher and the causal systems with fixed-points suggested by McCain and Turner. The unifying *augmented preferential semantics*, emerging as a result of this study, captures both Principles of Change and shows their clear and distinct roles.

Keywords reasoning about action · causality · preferential semantics · cognitive robotics

Mathematics Subject Classifications (2000) 68T27 · 06A06 · 03B44

1 Introduction

Reasoning about Action and Change is one of the most intriguing and fundamental issues in Artificial Intelligence. An intelligent agent is expected to interact with

M. Prokopenko (✉)
Information and Communication Technologies Centre, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Locked Bag 17,
North Ryde, NSW, 1670, Australia
e-mail: mikhail.prokopenko@csiro.au

its environment and reason about the interactions. Sometimes, the effects of an agent's actions can be traced relatively easily. On other occasions, an action may result in intricate and convoluted ramifications. Arguably, agents' ability to reason about direct and indirect effects of actions is a distinguishing feature of intelligence. Ultimately, agents' existence and survival in the environment depends on their competence in reasoning about changes in the environment.

Reasoning about actions and change may take many forms. For example, behaviour of simple biological organisms and basic situated artificial agents embeds reasoning about change in low level reactions. More complex life forms (natural or synthetic) are able to represent the environment, model it and reason about consequences of their actions. One fundamental characteristic of such representations is an *explicit* notion of change, or in other words, an incorporation of "time's arrow" (the temporal asymmetry). For example, an agent may consider that events depend on earlier events in a way in which they do not depend on later events, and subjectively deliberate for the future on the basis of information about the past.

It is well recognised that complex tasks such as prediction, planning, explanation assume some distinctions between the past, the present and the future and involve some form of temporal reasoning, including "reasoning about phenomena that take place in time, i.e., *reasoning about actions and change*" [34]. Whether an artificial agent is expected to calculate a moving object's position over time, determine the state of an electric circuit, or find out a reason for the fire that destroyed a house, it must assume (among others) some notion of change, action and causation. Ideally, if reasoning is expected to be consistent and systematic, these notions should be formalised, leading to reproducible and comparable results across agents. In other words, an intelligent agent may need a formal reasoning system that produces inferences about the effects of actions. However, a unique and completely general-purpose logic of reasoning about action and change is no longer perceived as the main research objective. It has been argued that "perhaps the logic of common-sense reasoning, rather than being unified and concise, will have the character of a Swiss army knife and contain one tool for each purpose" [34]. In other words, various reasoning systems may be based on different theories of action. This highlights the role of a underlying semantics that allows us to analyse and compare action theories.

One particular trait emerging in recent literature on Reasoning about Action attempts to explicitly embody a notion of causality (causation) in logic-based action theories. It is argued that such an extension would not only enhance the expressive power of theories of action, but may also provide more concise representations [11, 15, 18, 19, 40]. What seems to be lacking so far is a general semantic framework covering this particular class of action theories. The presented work attempts to examine some of the aspects of "time's arrow" and causality, explores the role of several important underlying principles, and introduces a general semantics for a class of action theories represented by a number of recent and influential approaches.

2 Preliminaries and background

The area of Reasoning about Action has grown considerably in the last decades, and overlaps now with many other fields, as diverse as philosophy of causation and robotic soccer. This can be partially explained by the fact that many related, though distinct, areas share some essential problems, crystallised and investigated within the field of Reasoning about Action (such as planning, explanation, prediction).

Following Sandewall and Shoham [34] we say that a reasoning task typically involves "(1) designation of certain actions which have been (will be, may be) performed, as well as their order of execution; (2) statements about the state of the world before the actions; (3) statements about the state of the world after the actions." In a planning task, (2) and (3) are given and (1) is sought, while a prediction task uses (1) and (2) in determining (3). A more general interpretation is the (extended) prediction problem, referred to by Shoham [37] as a problem of "how to reason *efficiently* about what is true over extended periods of time," while maintaining "certain tradeoffs between risk avoidance and economy." In short, the extended prediction problem is that an agent needs to make a lot of predictions about short future intervals before predicting something about the more distant future.

It is interesting to note that two challenging problems in Reasoning about Action – the *Frame* and *Ramification* problems – are related to (and arguably, can be subsumed by) the extended Prediction problem. Informally, the Frame problem is concerned with what does not change when an event occurs or an action is performed. Sometimes, the term "Frame problem" is given a broader scope, but typically it is used in the restricted sense of the *Persistence* problem: assuming that properties of the world do not change unless affected by an action (an event), the aim is to build a reasoning system that models the dynamics of the world in an efficient way. In most cases, however, it is not sufficient just to update directly affected properties and leave the rest unchanged – some action consequences may spread quite far and affect seemingly remote and unrelated properties (for instance, the "domino effect" scenario). In other words, the agent also faces the *Ramification* problem – how to formalise all of the things that do change as the result of an action. Ginsberg and Smith [9] describe the problem as follows:

> The difficulty is that it is unreasonable to explicitly record all of the consequences of an action, even the immediate ones. ... For any given action, there are essentially an infinite number of possible consequences that depend upon the details of the situation in which the action occurs.

It is precisely the combination of the Frame and Ramification problems that makes a search for a concise solution extremely challenging and non-trivial. Logic has traditionally been chosen as the representation language and various reasoning systems have been designed to address the combined problem: situation calculus [20], default logic [32], circumscription [13, 21], temporal logic of chronological ignorance [37, 38], action languages [10, 12, 14, 43], fluent calculus [40], features and fluents framework [33], the Possible Worlds Approach (PWA) [9], the Possible Models Approach (PMA) [45], causal relationships approach [39, 40], causal fixed-points [18, 19], event calculus [36], nonmonotonic causal theories [11], etc.

Usually, most solutions require that action specifications provide *direct* (most significant, immediate, etc.) *effects* explicitly, and employ domain constraints of some form for specifying additional (indirect) changes that may occur due to the action. The monotonic situation calculus and some non-monotonic logics try to *infer* which propositions are true once the actions have been performed, and answer queries about the theory without actually updating it. An alternative way to formalise reasoning about change was proposed in the STRIPS approach [6] and extended in the PWA and the PMA. “The basic insight . . . is that the world does not change much from one instant to the next” [9]. So it is possible to maintain a current state of the world and incorporate an update procedure constructing “the nearest world to the current one in which the consequences of the actions under consideration hold.” In other words, an agent follows the *Principle of Minimal Change*, reducing the amount of explicit information about what changes and what persists through an action. According to this principle, the world changes as little as *possible* when an action is performed. A precise definition of minimal change depends on the particular formalism in question. Often, it is defined by set inclusion, presuming that the total set of changes resulting from an action contains those changes that are explicitly specified as direct effects of the action, and a *minimal* set of other changes required by the domain constraints. Sometimes, a particular measure of minimal change assigns different degrees of inertia to properties under consideration (a policy of *categorisation*), which allows a reasoning system to assume persistence for more basic (independent) properties and apply domain constraints to secondary (derived) properties. In general, an agent uses a preference relation in accepting outcomes (states, sets of states, interpretations) that are strictly closer to the initial one than other possibilities (which are rejected).

In addition, some action theories embody background information in the form of domain “causal rules” or constraints, and apply the *Principle of Causal Change*. Informally, these approaches specify how changes in one property (state variable, state of affairs, event, state) may “cause” or influence changes in another. In response to an action, a reasoning system is expected to produce the outcome which satisfies the action’s direct effects and the domain constraints, while incorporating only changes justified by the underlying causal constraints.

Sometimes, the Principles of Minimal and Causal Change are applied together, resulting in policies of causal minimisation [18, 22]. In other words, an agent reasons that the world changes as little as *necessary* when an action is performed. One particular approach following this kind of causal minimisation is McCain and Turner’s approach [18] that introduces *causal fixed-points*. Intuitively, a causal fixed-point is an outcome incorporating the direct effects of actions, where all other changed properties are causally justified (in a certain sense). In other words, every detail in the outcome must be “explained” either as persisting through the action, or as a direct effect, or as a causal ramification of *other properties contained in the outcome* – hence the fixed-point flavour. Obviously, a possible outcome that contains at least one detail without such justifications is rejected, even if it does not violate the domain constraints. It is not hard to observe that causal fixed-points, indeed, incorporate only necessary changes.

It has been argued in some recent proposals that the Principle of Minimal Change can be replaced or subsumed by the Principle of Causal Change in reasoning about actions: “the aim of generating ramifications is not to minimize change but to avoid changes that are not caused, which . . . need not be identical” [40]. These approaches allow a reasoning system to propagate beyond just nearest possible states towards states where all changes are justified. One interesting example is the Light Detector example, proposed by Thielscher [40]. This example illustrates that sometimes one possible successor to an initial state of the world may have strictly more changes than another successor, while both of them seem to be intuitive. An electric circuit includes three switches, a relay, a light bulb and a light detector device (figure 1).

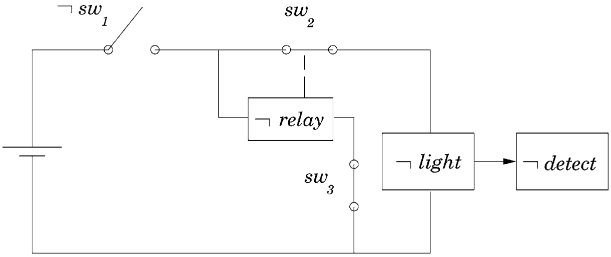


Figure 1 The electric circuit with light detector.

Initially, both the light bulb and the detector are off. The circuit is specified in such a way that it is possible (by toggling one of the switches) to activate both a relay and a sub-circuit involving the light bulb for an instant – before another switch jumps its position as a ramification of activating the relay, and turns the light off. It is argued by Thielscher [40] that at this brief instant the detector may react to the light. Despite the fact that the light bulb itself does not stay activated, the detector may. Therefore, two outcomes are presented as possible: one where the light is off, and the detector is not activated, and another, where the light is off as well, but the detector is activated. Obviously, the second outcome has strictly more changes (with respect to the initial state) than the first – the change in the detector's state is an extra ramification. Importantly, this change is justified "during" some dynamic process employing causality (e.g., propagating from "no light and no detector" to "light and no detector" to "light and detector" to "no light and detector"). Unlike fixed-point details, the presence of the activated detector is justified not by other "contemporary" properties (statically) contained in the outcome, but rather by some (dynamic) propagation of change.

The *causal relationship approach* described by Thielscher [40] formalises a proposal capturing both successors in this example, and argues that the Principle of Minimal Change "is not always adequate for distinguishing between possible indirect effects on one hand, and unfounded changes on the other hand." However, it appears to be extremely hard to compare the roles of the two principles of change *within* particular action theories. Not only are there different interpretations across the field, but the principles' manifestations are often limited by the operational mechanics of particular reasoning approaches. This work aims to investigate the common ground taken by different approaches to reasoning about action and change. In Section 3, we shall attempt to set a framework for a general semantics, relying on the Principle of Minimal Change and the Principle of Causal Change, and clarify the reasons that allow us to hope that our motivating approaches can be represented in a unifying setting.

3 Framework

An agent reasoning about action and change may represent a dynamic world in many ways, choosing certain components and discarding others. In this section we shall discuss different aspects of world dynamics and its representations, while trying to develop our framework incrementally.

3.1 States and actions

A general semantics for a class of action theories should not rely on the choice of specification languages, and "can only be achieved if we model properties without referring to the internal structure of world states" [26]. More precisely, we shall denote the set of all world states defined for a specific representation scheme[1] by $\mathcal{W}$, and consider a world state as an uninterpreted point in the space $\mathcal{W}$. We intend to demonstrate that most of the crucial concepts can be captured in this representation-independent style.

How can an agent represent change? What are the aspects of temporal and causal asymmetries that the agent may perceive, represent and reason about? These questions lie at the very core of Reasoning about Action, and not surprisingly, may be answered from very different philosophical viewpoints: "almost *any* change can be thought of both historically (in terms of sequence of states, i.e., a change of state) and experimentally (as a new *kind* of state, a state of change)" [7]. According to the state-based approach, "a state is a snapshot of the underlying dynamic system, i.e., the part of the world being modeled, at a particular instant of time" [40]. It could be argued that "change arises as a by-product of the assignment of states to times" [7], and the history of the world is a (temporally) ordered set of states. An alternative, event-based account of change, recognises only those states which can be characterised in terms of events. In a *mixed* account, however, both states and events are admitted as primitive terms, and "one has to specify the logical and causal relationships which hold between states and events" [7]. Following the mixed account of change, we introduce a finite set of events (or actions) $\mathcal{E}$, without speculating about the internal structure or type of the events.

We now need to indicate how an action's effects can be reflected in the state-space $\mathcal{W}$. A natural way is to introduce the post-condition of an action $e \in \mathcal{E}$ as the property that e directly brings about with its occurrence. We shall denote the post-conditions of the action e by $[e]$, defined as a subset of $\mathcal{W}$. Intuitively, the post-condition of e is precisely the properties common to all states in $[e]$. The post-conditions $[e]$ are not made conditional on the initial states where the action may be executed, and therefore, are captured unvaryingly and uniformly by a subset of $\mathcal{W}$. Whenever the action e is performed, an agent considers states in the set $[e]$ as states compatible with the action's direct effects. Formally, we define $[e]$ to be a function from $\mathcal{E}$ to $2^{\mathcal{W}}$ (the power-set of $\mathcal{W}$), such that for all actions e in $\mathcal{E}$,

$$[e] \subseteq \mathcal{W}, \quad [e] \neq \emptyset.$$

In other words, every action is satisfied by at least one state.

3.2 Possible, legitimate and successor states

The state-space $\mathcal{W}$ contains, in principle, all *conceivable* states of an underlying dynamic world (system). Given a particular representation scheme and a choice of state variables, one may consider each combination of variable components as a possible state, and together they make up the conceivable state-space. This view

[1] For our purposes, it is sufficient to consider this set to be a finite set.

makes use of the idea of a model of the world that satisfies the requirements of *Logical Atomism* [44, p. 110]. There is a set of n basic features (or states of affairs), and a state of the world, at any given time, is a conjunction with n terms such that each of the basic features or its negation appears as a term. Hence, there are 2^n states that are logically possible. Given a sequence of k events (actions) or "occasions" following von Wright's terminology [44, p. 108], forcing state transitions, the number of all possible successions (histories) of the world is 2^{kn}. However, since the basic components may be inter-related and mutually restricted, not every combination represents a nomologically (lawfully) possible state. This conjecture definitely presupposes the existence of underlying laws [2]:

> Only those values of the components of the total state function that are compatible with the laws will be really (not just conceptually) possible. In other words, because the laws impose restrictions upon the state functions and their values, hence upon the state spaces, only certain subsets of the latter are accessible to the thing represented. We shall call the accessible part of the state-space the *lawful state space* of the thing in the given representation and relative to a given frame.

In short, the lawful state-space is a *proper* subset of the state-space $\mathcal{W}$. The elements of this subset (lawful states) are sometimes referred to as *admitted* [35] or *legitimate* [26] states. In the context of many logics of action, legitimate world states are defined as the elements of $\mathcal{W}$ satisfying certain conditions known as *domain constraints*. The domain constraints are often specified through state variables and syntax-dependent relations, and therefore, shall not be used directly in our representation-independent approach.

Instead, we introduce the set $\mathcal{D}$ of legitimate states explicitly – as a proper non-empty subset of $\mathcal{W}$. Ideally, given a particular scheme representing domain constraints, our semantics will identify the elements of the set $\mathcal{D}$ with the states satisfying these constraints. What is needed now is a concise way to specify the states resulting from an action $e \in \mathcal{E}$ executed at an initial state $w \in \mathcal{W}$ (or more precisely, at a state $w \in \mathcal{D}$, excluding the possibility of being in an illegitimate state in the first place). In other words, we need to describe the *successor* state(s), where the underlying system moves to as a result of an action e performed at a state w. Formally, we define a *selection* or *result* function $Res(w, e)$ to be a function from $\mathcal{W} \times \mathcal{E}$ to $2^{\mathcal{W}}$, mapping a state w and an action e to a set of (legitimate) successor states. An example of the selection function is a simple function choosing all legitimate states compatible with the action's direct effects:

$$Res_*(w, e) = \mathcal{D} \cap [e].$$

Obviously, this particular function would not, normally, satisfy an intelligent agent, since the selected successor states have no connection to the initial state w (shown as the left-most circle in figure 2). In general, we do not require that a successor state necessarily satisfies the action's direct effects, i.e., for any action e and state w,

$$Res(w, e) \subseteq [e].$$

Although this is a very sensible assumption, it was argued in some proposals that a successor state may be a result merely *triggered* by an action's post-conditions [40].

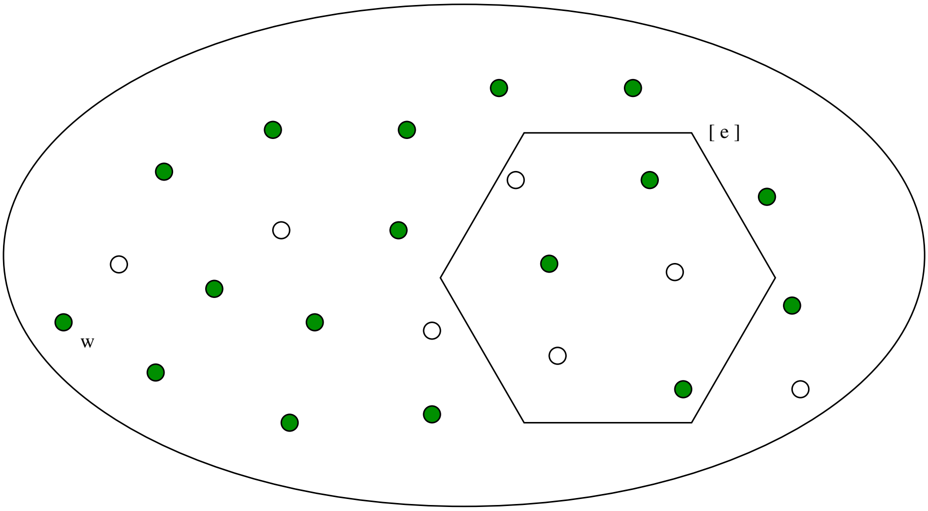


Figure 2 The legitimate states (members of the set $\mathcal{D}$) are depicted as filled circles. The states compatible with direct effects of the action e are enclosed within the polygon area $[e]$, and the simple selection function $Res_*(w, e)$ chooses all three states depicted as filled circles, inside the polygon.

We shall refer to the view that a successor state must satisfy the direct effects as *conservative*, and highlight this distinction in further analysis. In addition, we do not make another intuitively appealing assumption that for any action e,

$$\mathcal{D} \cap [e] \neq \emptyset.$$

This means that we do not impose the requirement that, for every action, there must always be a state where the action post-conditions co-exist with domain constraints. This requirement may, however, become quite reasonable if the conservative view on successor states is accepted.

Following [27], we shall say that an action theory with a function $Res_1(w, e)$ is *selection-equivalent* to an action theory with a function $Res_2(w, e)$ if and only if $Res_1(w, e) = Res_2(w, e)$, for every action e and state w. A general unifying semantics would support translations between action systems which use different definitions of selection functions.

3.3 Information states

It is well known that one may distinguish two distinct transitions: the transition between the world states r_1 and r_2, brought about by an event e, and the transition between the agent *information* states γ_1 and γ_2, triggered by the perception of the event e or execution of the action e. For example, Peppas [26] distinguished between a general process called *system dynamics* described as “the process by which a dynamic world changes states due to the occurrence of the events,” and another general process called *knowledge dynamics* explained as “the process by which an agent changes beliefs about the current world state, in the light of information about the occurrence of an event.” While this distinction has been identified, the information states were not included in the formal consideration. Instead, the difference was related to different forms of belief change – revision and update. We suggest that a better way to capture this fundamental distinction is to explicitly introduce the set of all information states, denoted as Γ.

Intuitively, an information state is a state of an agent's reasoning process. While reasoning, the agent may consider previous and current states of the external world, contemplate potential histories of state transitions, contextualise causal knowledge, and so on. All these rather partial information sources contribute to the reasoning process and fuse into more comprehensive information states. Therefore, in a typical case, an information state has more dimensions than a state of the external world entertained by the agent. We do not intend here to associate the notion of *information states* with any particular neuro-biological concept, such as conscious, mental, or brain states. What is important is the distinction between dynamics in the information state-space Γ, and system dynamics in the state-space $\mathcal{W}$.

This approach does not necessarily question the view of logical atomism, or the position that a world state can be completely described by all relevant facts about it, or the agents' ability to reason about both conceivable and nomologically possible world states. Rather, we extend these views by allowing the agent's reasoning process to use information states with entirely different dimensions. Most importantly, an information state does not have to be uniquely associated with a time reference. Intuitively, the agent, motivated by a single action (event) e, may process a whole series of state transitions in the information state-space Γ before making a judgement on the world transition from $w \in \mathcal{W}$ to one of the successor states in $Res(w, e)$. In short, information states are *not* the agent's beliefs about the state of the world, but *points in the information space through which the agent's reasoning may navigate*.

Technically, some action domains may avoid the distinction, resulting in equating the world and information state-spaces: $\mathcal{W} = \Gamma$. In other words, each information state entertained by an agent corresponds exactly to one world state. This approximation may well be the reason for fusing the information state-space with the external world state-space in many various approaches. Some recent proposals discern the difference by employing concepts of hyper-states [28] and power-states [29], but without giving a clear intuition on the nature of these additional concepts. We will demonstrate that some action theories may be captured by our semantics while staying with the approximation $\mathcal{W} = \Gamma$, whereas others require $\mathcal{W} \neq \Gamma$.

Now we can introduce a *projection* function $\mathcal{P}$ from Γ to $\mathcal{W}$ – this function maps an information state $\gamma \in \Gamma$ to a world state $w \in \mathcal{W}$. Intuitively, the projection function "extracts" the world state "component" from a more convoluted information state. We require that for any state $w \in \mathcal{W}$ there exists an information state $\gamma \in \Gamma$ such that $\mathcal{P}(\gamma) = w$. In other words, the function $\mathcal{P} : \Gamma \to \mathcal{W}$ is a surjection (i.e., the function's image is its codomain, and $\mathcal{P}$ can return any value in $\mathcal{W}$).

We also derive a *set-projection* function $\mathcal{X}$ mapping sets of information states onto sets of world states from $\mathcal{W}$. The function $\mathcal{X} : 2^{\Gamma} \to 2^{\mathcal{W}}$ is defined as follows: $\mathcal{X}(\{\gamma_1, ..., \gamma_n\}) = \{\mathcal{P}(\gamma_1), ..., \mathcal{P}(\gamma_n)\}$. It is clear that for any set $\Pi \subseteq \Gamma$, we obtain that

$$\mathcal{W} = \mathcal{X}(\Pi) \cup \mathcal{X}(\Gamma \setminus \Pi) = \mathcal{X}(\Gamma),$$

although, in general,

$$\mathcal{X}(\Pi) \cap \mathcal{X}(\Gamma \setminus \Pi) \neq \emptyset.$$

In addition, we define a set of information states $[e]^\Gamma$ as $\{\gamma \in \Gamma : \mathcal{P}(\gamma) \in [e]\}$. In other words, $[e]^\Gamma$ denotes the set of information states whose world state-space projections are contained in the set $[e]$. By definition, $\mathcal{X}([e]^\Gamma) = [e]$. The following abbreviation is also useful: $\mathcal{D}^\Gamma = \{\gamma \in \Gamma : \mathcal{P}(\gamma) \in \mathcal{D}\}$. The set $\mathcal{D}^\Gamma$ contains all the information states that would project to the legitimate states in $\mathcal{D}$. Again, by definition, $\mathcal{X}(\mathcal{D}^\Gamma) = \mathcal{D}$. The information states outside $\mathcal{D}^\Gamma$ are not illegitimate information states per se – the notion of being legitimate applies only to the states in $\mathcal{W}$. In fact, the information states outside $\mathcal{D}^\Gamma$ are quite acceptable and useful in the reasoning process, providing important intermediate steps for state transitions.

Figure 3 illustrates the information state-space Γ (the top part of the figure), while showing the states in $\mathcal{D}^\Gamma$ as filled circles, and enclosing elements of $[e]^\Gamma$ in the polygon. The projection $\mathcal{P}$ of some information states onto the state-space $\mathcal{W}$ (the bottom part of the figure) is shown with arrows.

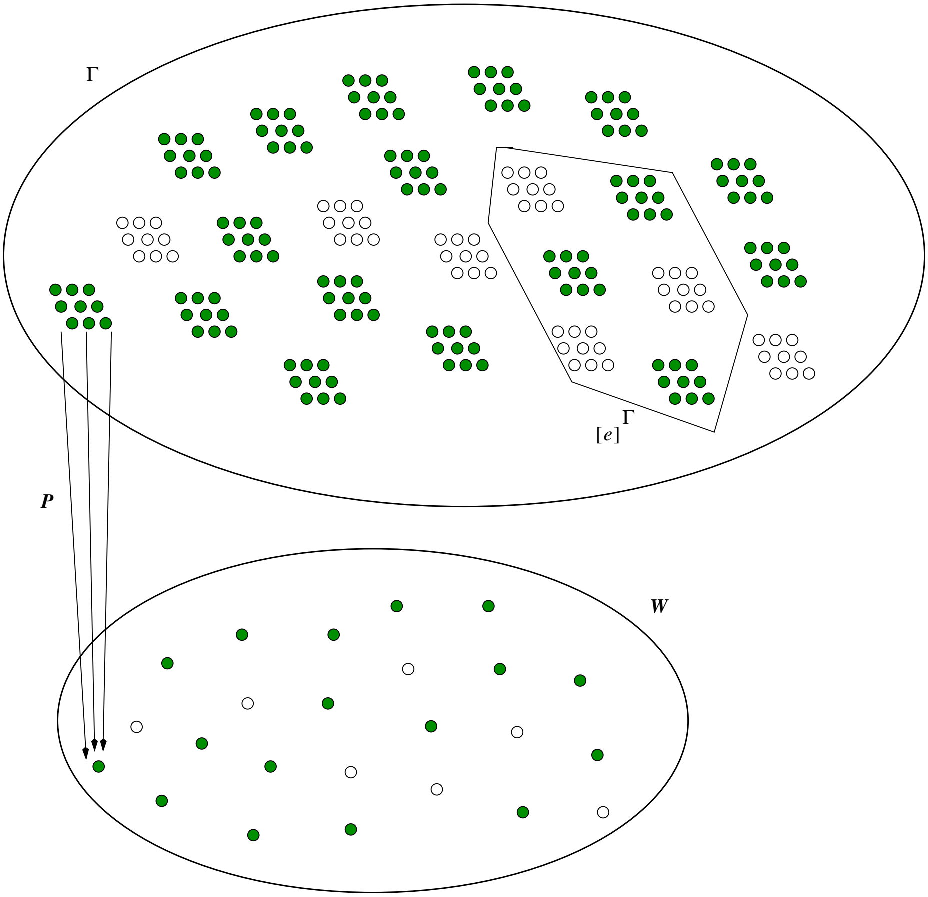


Figure 3 The information state-space Γ and the projection function $\mathcal{P}$.

Let us illustrate how these notions can be used in the selection function. Consider again the simple function choosing all legitimate states compatible with an action's direct effects. Now, we can represent this in terms of information states as

$$Res^*(w,e) = \mathcal{X}(\mathcal{D}^\Gamma \cap [e]^\Gamma) = \mathcal{D} \cap [e] = Res_*(w,e).$$

In other words, this function selects projections of those information states that are common to both sets $\mathcal{D}^\Gamma$ and $[e]^\Gamma$.

3.4 Preference relation on states

The selection function $Res^*(w,e)$ choosing all legitimate states compatible with an action's direct effects produces successor states which are not related to the initial state w. It is hard to imagine that, in general, an action results in a state which may be arbitrarily different form the initial one. One way to address this question is to assume the existence of inertia and apply the Principle of Minimal Change:

> whenever an event e occurs at some world state w, a successor state $r \in Res(w,e)$ must satisfy the post-condition $[e]$ and differ as little as possible from w with respect to some (local) measure of change.

Constructively, there exists an ordering on states $<_w$ reflecting the comparative degree of change between w and a potential successor state. This allows an agent to judge that $r_1 <_w r_2$ if and only if the degree of change between w and r_2 is at least as great[2] as the one between w and r_1. An agent normally maintains a different ordering $<_w$ for each state w, resulting in a set of orderings: a preferential structure.

The question that naturally arises now is whether we wish to consider a preferential structure only on world states in $\mathcal{W}$ or on information states in Γ as well. This would not, of course, identify minimal change in the world with minimal change in the agent's beliefs about the world[3] – simply because information states are different from the beliefs and may be entitled to their own preferential structure. Besides, a preferential ordering on information states would not introduce syntax-dependent measures of change. Therefore, it is quite permissible, we believe, to consider a set of orderings on world states or a set of orderings on information states – either method captures the Principle of Minimal Change in its own way. It also appears that one should start with a weaker assumption that only one such set is needed in a general semantics. Normally, it should be possible to derive one preferential structure from another: for a projection function $\mathcal{P}$, and a given specification of an ordering $<_w$ (for each w) defined on $\mathcal{W} \times \mathcal{W}$, one may produce an ordering $\ll_\gamma$ (for each γ) defined on $\Gamma \times \Gamma$, and vice versa.

We choose to operate with a preferential structure $\mathcal{O} = \{<_\gamma : \gamma \in \Gamma\}$ defined on information states, without claiming that it is richer or more intuitive than a structure on world-states. In fact, when we discuss an action theory that needs the distinction $\mathcal{W} \neq \Gamma$ (Section 7.1), we derive preferential orderings for information states $\ll_\gamma$ out of simple orderings $<_w$ defined for world states.

[2] For simplicity, we shall use the notation $<$ for a non-strict ordering.

[3] This was, in fact, the main reason for a failure of the Possible Worlds Approach [9], as shown by Winslett [45].

One interesting ordering type is the so-called PMA ordering, based on the Possible Models Approach [45]. In order to describe this ordering constructively, we need to assume certain internal structure for states. For instance, let us consider n basic truth-valued fluents, and let each state be a set with n elements such that each of the basic features or its negation appears as an element. We also define the symmetric difference between two states x and y to be the set $Diff(x, y) = (x \setminus y) \cup (y \setminus x)$, where $x \setminus y$ denotes set subtraction. For example, if $r = \{a, b, c\}$, $p = \{a, \neg b, c\}$, and $q = \{a, \neg b, \neg c\}$, we obtain $Diff(r, p) = \{b, \neg b\}$ and $Diff(r, q) = \{b, c, \neg b, \neg c\}$. We shall say that a state y is preferred to a state z relative to x in terms of the PMA ordering $\prec_x$, denoted $y \prec_x z$, if and only if $Diff(x, y) \subseteq Diff(x, z)$. Intuitively, it means that state y differs less from x than the state z does from x in terms of basic features. Continuing the example with three states r, p and q, we immediately obtain that $p \prec_r q$. Figure 4 depicts direct PMA preferences for eight states definable with three truth-valued fluents a, b and c, with respect to the state $r = \{a, b, c\}$ (more distant states appear further to the right from r). Note that states in each vertical layer are not mutually comparable in terms of the PMA ordering – PMA ordering is not total. Figure 5 shows a subset of the original ordering, such that states from different areas enclosed by pseudo-concentric dashed curves are PMA-comparable.

The concept of a preferential structure $\mathcal{O} = \{<_\gamma : \gamma \in \Gamma\}$ is inspired by Shoham's *preferential semantics* [38] for a class of non-monotonic logics. Under this idea an ordering is placed over the class of interpretations. The models corresponding to a particular inference are then identified as the minimal models under this ordering that satisfy the premises: only the most preferred (most plausible) interpretations are tolerated as serious possibilities. In terms of our framework, the preferential structure $\mathcal{O}$ suggests a very intuitive way to define a selection function $Res(w, e)$. Specifically, for a simple case approximating $\mathcal{W} = \Gamma$, the function $Res(w, e)$ should choose those legitimate states compatible with the post-condition $[e]$ that are nearest to w in terms of $<_w$. More precisely,

$$Res_1(w, e) = min(<_w, \mathcal{D} \cap [e]).$$

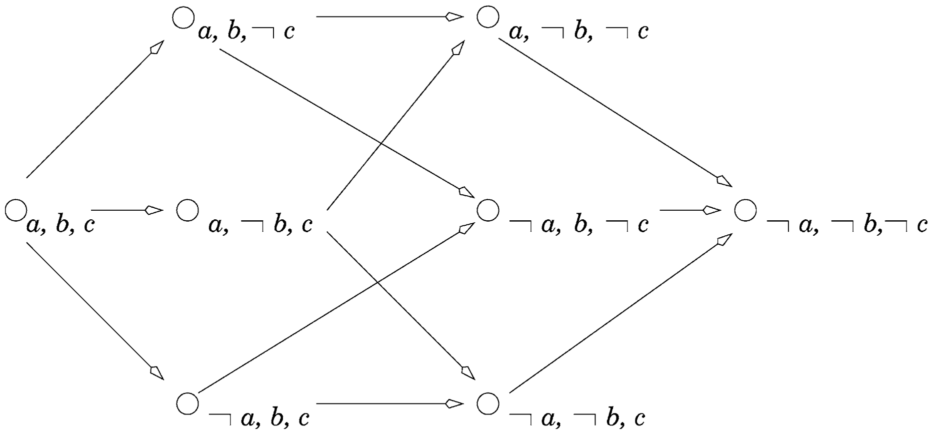


Figure 4 A partial PMA ordering: arrows point to more distant states.

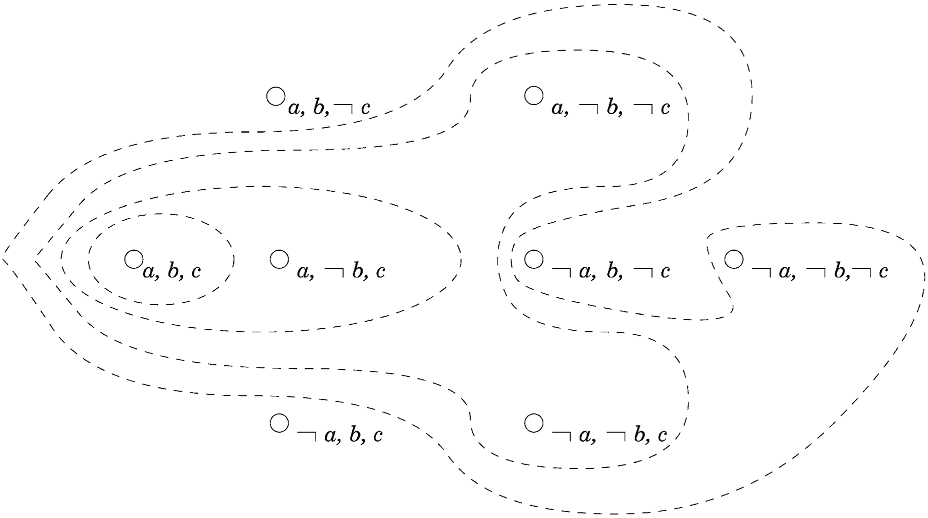
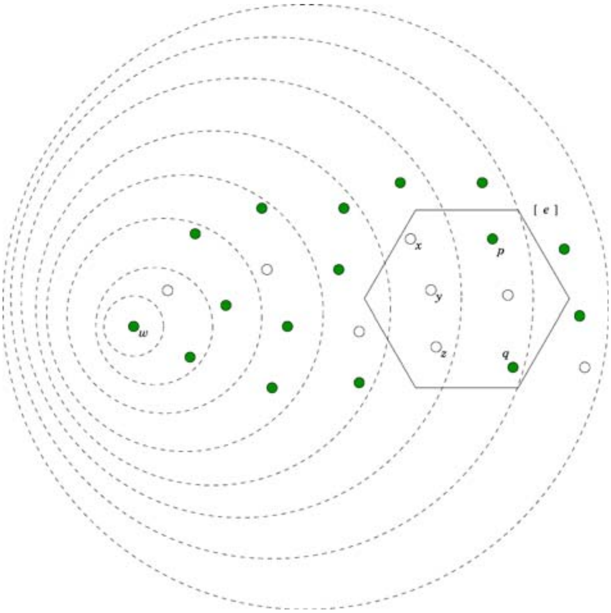


Figure 5 A subset of PMA ordering: "expanding" boundaries of minimality.

Figure 6 illustrates this selection (again, more distant states, enclosed in the expanding minimality-driven circles, appear further to the right from the left-most initial state w). There exists a stronger version of the preferential selection function:

$$Res_0(w, e) = min(<_w, [e]) \cap \mathcal{D}.$$

Figure 6 The selection function $Res_1(w, e)$ contains two states p and q, while the selection function $Res_0(w, e) = \emptyset$, given that the set $min(<_w, [e]) = \{x, y, z\}$ contains no legitimate states.



Here, an agent first selects the $<_w$-minimal states among the post-condition states $[e]$, and then chooses legitimate states out of the selection, if there are any (figure 6). If, for example, the state y was legitimate, then it would be (uniquely) selected by $Res_0(w,e)$ as well as $Res_1(w,e)$. Clearly, for any state w and action e,

$$Res_0(w,e) \subseteq Res_1(w,e) \subseteq Res_*(w,e).$$

We assumed here an approximation $\mathcal{W} = \Gamma$, and used a preferential structure over world-states. Similar preferential selection functions can be defined for the case when $\mathcal{W} \neq \Gamma$. Let an information state $\alpha(w)$ be such that $\mathcal{P}(\alpha(w)) = w$ for a state $w \in \mathcal{W}$. We also assume that all information states $\alpha(w)$ which project onto the same world state are assigned the same ordering $<_{\alpha(w)}$. Then we may specify the following selection functions.

$$Res^1(w,e) = \mathcal{X}(min(<_{\alpha(w)}, \mathcal{D}^\Gamma \cap [e]^\Gamma)).$$

and

$$Res^0(w,e) = \mathcal{X}(min(<_{\alpha(w)}, [e]^\Gamma) \cap \mathcal{D}^\Gamma).$$

Here, the agent selects minimal states in the information state-space by using the preferential structure $\mathcal{O}$ on information states, and then projects them onto successor world states. It is also obvious that, for any state w and action e,

$$Res^0(w,e) \subseteq Res^1(w,e) \subseteq Res^*(w,e).$$

A minimisation in the information state-space Γ followed by a projection onto world state-space $\mathcal{W}$ may result in different outcomes compared with a direct minimisation in $\mathcal{W}$ – but neither domain of minimisation ($\mathcal{W}$ or Γ) leads to stronger successor selections in general. Another important aspect is that, regardless of the domain of the preferential structure, each ordering $<_w$ is dependent only on the initial state w, and is not contingent on an action. Otherwise, cumbersome preferential structures containing orderings for each state-action pair $<_{w,e}$ would seriously undermine the quest for conciseness, given the number of potential combinations and weak elaboration tolerance.

In summary, the steps we have taken so far towards a general semantics, can be reflected in a tuple $\langle \mathcal{W}, \ \mathcal{D}, \ \Gamma, \ \mathcal{E}, \ \mathcal{O} \rangle$ and the functions $[e] : \mathcal{E} \to 2^{\mathcal{W}}$ (post-conditions), $\mathcal{P}(\gamma) : \Gamma \to \mathcal{W}$ (projection), and $Res(w,e) : \mathcal{W} \times \mathcal{E} \to 2^{\mathcal{W}}$ (selection).

3.5 Causal rules, causal relations, and causal laws

Despite many intuitive features of preferential style semantics, it has been shown recently that sometimes there are additional forces "producing" successor states. The main explanation suggested in the literature so far is that domain constrains (reflected in the legitimate states $\mathcal{D}$) are not sufficient for capturing all the domain dependencies, and are not strong enough to properly trace the ramifications of actions. Therefore one needs to explicitly introduce *causal* constraints and use the force of causality in complementing minimality. For example, McCain and Turner's causal theory of action [18] uses "causal laws," Thielscher's [40] framework is based on "causal relationships," Sandewall [35] proposed a causal relation on states in order to capture ramifications cascading to successor states, and Lin [15] extended the Situation Calculus with a ternary predicate $Caused(p, v, s)$ set to be true if the proposition p is caused by something unspecified to have the truth value v in the state s. In the following sub-sections, we would like to briefly review some terminology related to causally driven state transitions.

3.5.1 Causal rules – ontological dimension of causation

Two sorts of causation subjects were identified by Mellor [23]: *states of affairs* ("sentences, statements or propositions"), and *particulars* (things or events). This duality is well recognised in Reasoning about Action. In specifying the effects of actions, Lin [15] distinguished between

- *Action-triggered* causal statements (e.g., the action *load* causes the gun to be *loaded*) and
- *Fluent-triggered* causal statements (e.g., the fact that the *switch* is in the up position causes the *light* to be on),

and convincingly argued that *action-triggered* causation is convenient for representing direct effects of actions, while *fluent-triggered* causation – for indirect effects, or ramifications. Similarly, action languages (e.g., [14, 43]) differentiate between *static* and *dynamic causal laws*. A static causal law $\varphi \ causes_f \ \psi$ captures "fact causality" as a dependency between two facts contained in the same state. It works as an inference rule $\varphi \Rightarrow \psi$, expressing a causal "determination relation" between φ and ψ, and making the fluent formula ψ true when the fluent formula φ is true [18, 43]. A dynamic causal law $A \ causes \ \varphi \ if \ \psi$ expresses "event causality" by specifying which changes are caused by performing an action A [14]. Another classification is given by Geffner [8], who discriminated between *non-temporal* and *temporal causal rules*. A non-temporal rule says that an effect is true in the same state where its cause is true, while a temporal rule says that an effect is true in the state *following* the state where its cause is true. In short, the causal connections among states of affairs are captured by fluent-triggered causality, and the causal connections between particulars are similar to action-triggered causality.

3.5.2 Causal relation – epistemological dimension of causation

The attempts of preferential-style selection functions $Res_1(w, e)$ and $Res^1(w, e)$ defined in Section 3.4 are not entirely adequate because minimal legitimate states among post-condition states $[e]$ do not necessarily reflect fluent-triggered causality. While all the states in $[e]$ agree on the post-conditions of the action in question, accounting for action-triggered causality, the selection of minimal states may often be insufficient for capturing various causal dependencies. Thus we intend to augment our framework with a component explicitly targeting fluent-triggered causality. Instead of committing to a formal logical language and specifying fluent-triggered causal statements in some syntactic form via fluents or basic "states of affairs," it is possible to capture the underlying constraints in a causal binary relation on states. This choice manifests another dual aspect of causation – it may be expressed both intensionally (with a causal relation on fluents) and extensionally (with a causal relation on states). This duality is "orthogonal" to the event-fact distinction, and both types of causation (event driven and fact driven) may be represented with and without references to the internal structure of states (figure 7). For example, the causal relation *causes* can refer to (internal) state variables and/or events, while the causal relation "–>" can be (extensionally) defined on sets of states. Here, the notation $X \text{ –> } X \cap Y$ may indicate that the agent's reasoning process propagates from the states in the set X to the states in the set $X \cap Y$.

Consequently, in order to capture fluent-triggered causality extensionally, we introduce a binary relation $\mathcal{M}$ on information states. Two aspects require some explanation: the choice of the domain of the relation $\mathcal{M}$, and its properties. Is causation an *ontological* category? Is it a purely *epistemological* category, a theoretical relation, which belongs exclusively to our account of experience? A positive answer to the latter question, taken by empiricism, argues that "the status of causation category is purely epistemological, that is, causation concerns solely our experience with and knowledge of things, without being a trait of the things themselves" [1, p. 5]. According to Bunge [1, p. 6] who criticised the empirical doctrine,

> ...causation is not a category of *relation* among *ideas*, but a *category of connection and determination* corresponding to an actual trait of the factual (external and internal) world, so that it has an ontological status — although, like every other ontological category, it raises epistemological problems. Causation, as here understood, is not only a component of experience but also an objective form of interdependence obtaining, though only approximately, among real events, i.e., among happenings in nature and society.

The choice of the information state-space Γ (and not the state-space $\mathcal{W}$) as the domain for the relation $\mathcal{M}$ does not commit us to either view. The argument that

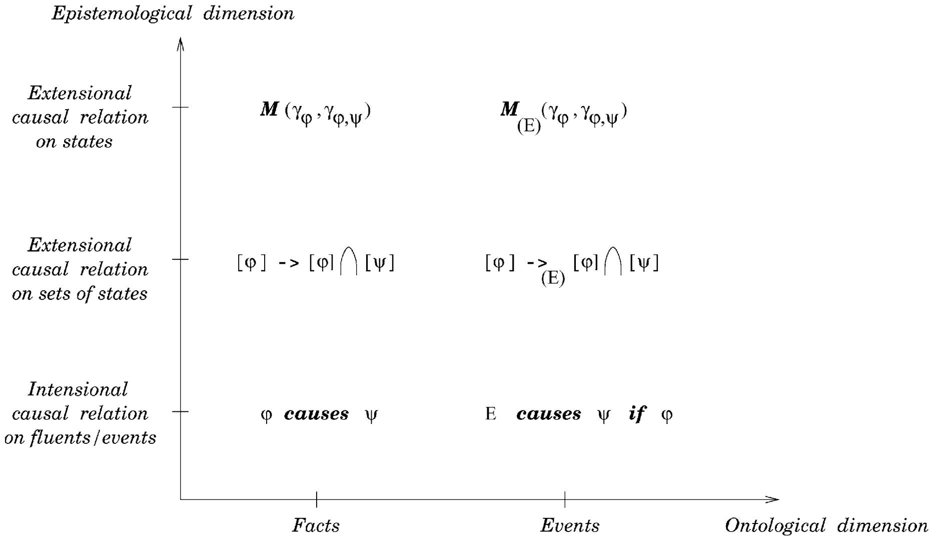

Epistemological dimension
Extensional causal relation on states
$\boldsymbol{M}(\gamma_\varphi, \gamma_{\varphi,\psi})$
$\boldsymbol{M}_{(E)}(\gamma_\varphi, \gamma_{\varphi,\psi})$
Extensional causal relation on sets of states
$[\varphi] \text{ -> } [\varphi] \cap [\psi]$
$[\varphi] \text{ ->}_{(E)} [\varphi] \cap [\psi]$
Intensional causal relation on fluents / events
φ ***causes*** ψ
E ***causes*** ψ ***if*** φ
Facts
Events
Ontological dimension

Figure 7 Ontological and epistemological dimensions of causation. Causal relations can be defined on states of affairs (fluents, events), on sets of states, on (information) states, etc.

causation is "an objective form of interdependence obtaining ... among real events," admits that causation may manifest itself "only approximately." Therefore, defining the causal relation $\mathcal{M}$ in terms of information states allows us to better approximate instances of interdependence between "real events." We believe that it is permissible to express the interdependence $\Rightarrow$ between two real states of affairs (or events) φ and ψ via a relation $\mathcal{M}_{(\Rightarrow)}$ between relevant information states γ_φ and $\gamma_{\varphi,\psi}$. If, on the other hand, causation does not apply to things but to experience alone, and is nothing but a direction enabling us to order or to label phenomena, then there is no harm in defining the *theoretical* causal relation in the information state-space and leaving functional, quantum and other objective dependencies to the world state-space.

Bunge observed [1, p. 244] that *the logical aspect of the causal problem is semantical rather than syntactical*. Following this view, we attempt to specify "the topology of the series, not the nature of its terms," while specifying formal properties of the binary relation $\mathcal{M}$:

($\mathcal{M}_1$) Irreflexivity: $\neg\mathcal{M}(\alpha, \alpha)$.

($\mathcal{M}_2$) Asymmetry: if $\mathcal{M}(\alpha, \beta)$ then $\neg\mathcal{M}(\beta, \alpha)$.

($\mathcal{M}_3$) Transitivity: if $\mathcal{M}(\alpha, \beta)$ and $\mathcal{M}(\beta, \gamma)$ then $\mathcal{M}(\alpha, \gamma)$.

The first property is straightforward: *nihil est causa sui* and, together with Asymmetry asserts the directionality of causation. Transitivity, however, is a more debatable property – we prefer not to require it and use, instead, the transitive closure of the relation $\mathcal{M}$, denoted $\mathcal{M}^*$, when needed. In other words, we rely on the propagation in the information state-space driven by the relation $\mathcal{M}$ without postulating the "cause-effect" relations between states not related in $\mathcal{M}$.

The binary relation $\mathcal{M}$ defined on $\Gamma \times \Gamma$ creates a multiplicity of causal chains $\mathcal{M}(\gamma_1, \gamma_2)$, $\mathcal{M}(\gamma_2, \gamma_3)$, ..., $\mathcal{M}(\gamma_i, \gamma_{i+1})$, ..., $\mathcal{M}(\gamma_{k-1}, \gamma_k)$, and therefore enables a propagation in the information state-space along them. Intuitively, such a transition process propagates some initial *change* towards an information state γ_k that is *stable* in terms of $\mathcal{M}$. In other words, the propagation stops when there are no possible causal links from γ_k. In short, the most essential role of the binary relation $\mathcal{M}$ is to provide some "topology" for tracing causal ramifications in Γ, in addition to ruling out non-stable states. Let us denote by $\mathcal{K}_\mathcal{M}$ the set of stable information states $\{p \in \Gamma : \neg\exists q \in \Gamma,\ \mathcal{M}(p, q)\}$. We also require

($\mathcal{M}_\mathcal{D}$) Density: $\mathcal{D} \cap \mathcal{X}(\Gamma \setminus \mathcal{K}_\mathcal{M}) = \emptyset$,

specifying that no unstable information state may be projected onto a legitimate state. The density condition implies

$$\mathcal{D} \subseteq \mathcal{X}(\mathcal{K}_\mathcal{M}).$$

Although a stable information state may be projected onto an illegitimate world state, a legitimate state is always a projection of a stable information state. The domain constraints reflected in the set $\mathcal{D}$ correspond to non-causal laws and may eliminate more illegitimate states than the causal relation alone – the elements of

$\mathcal{X}(\mathcal{K}_{\mathcal{M}})$ are just those states that do not conflict with causation in a given domain, from the agent's point of view. If $\mathcal{W} = \Gamma$, the original density condition reduces to

$$\mathcal{D} \subseteq \mathcal{K}_{\mathcal{M}}.$$

3.5.3 Causal laws – nomological dimension of causation

In this section we briefly consider the (existence of) underlying causal laws and their possible connections with causal relations, highlighting another dimension of causation – nomological (figure 8). Causal laws are identified with type-causality, capturing certain sorts of *regularity* or *recurrence*. It is unclear, however, whether all individual causal relationships are instances of universal causal laws: as noted by Brother William, in Umberto Eco's "The Name of the Rose" [5, p. 206],

> ... if only the sense of the individual is just, the proposition that identical causes have identical effects is difficult to prove. A single body can be cold or hot, sweet or bitter, wet or dry, in one place — and not in another place. How can I discover the universal bond that orders all things if I cannot lift a finger without creating an infinity of new entities? For with such a movement all the relations of position between my finger and all other objects change. The relations are the ways in which my mind perceives the connections between single entities, but what is the guarantee that this is universal and stable?

Indeed, this topic "has sown doubts" not only in the learned Franciscan's mind, but also in minds of many prominent philosophers. According to the reductionist

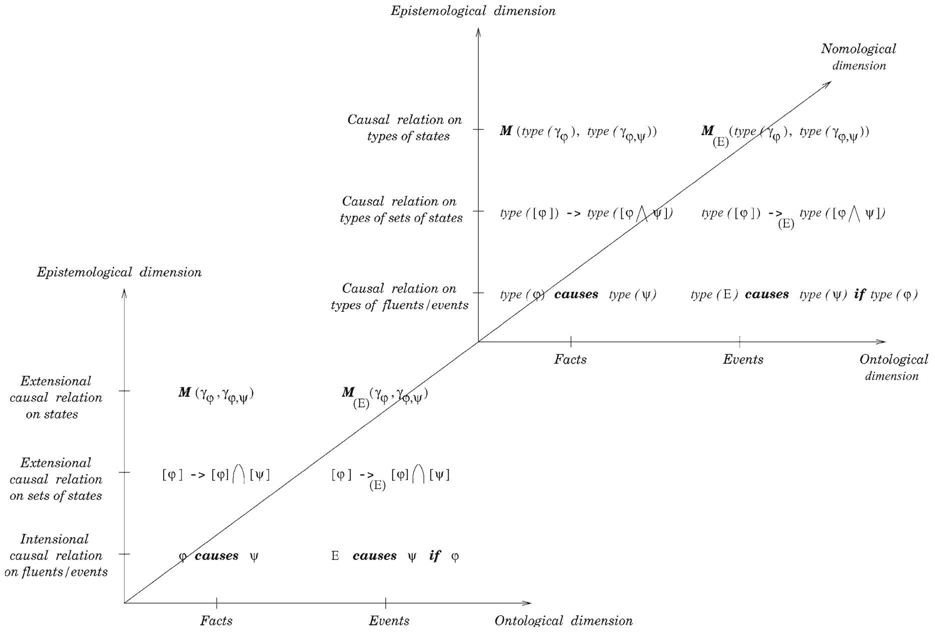

Epistemological dimension
Nomological dimension
Causal relation on types of states
$\boldsymbol{M}$ (type (γ_{φ}), type ($\gamma_{\varphi,\psi}$))
$\boldsymbol{M}_{(E)}$(type (γ_{φ}), type ($\gamma_{\varphi,\psi}$))
Causal relation on types of sets of states
type ([φ]) -> type ([$\varphi \bigwedge \psi$])
type ([φ]) $->_{(E)}$ type ([$\varphi \bigwedge \psi$])
Causal relation on types of fluents/events
type (φ) **causes** type (ψ)
type (E) **causes** type (ψ) **if** type (φ)
Facts
Events
Ontological dimension
Epistemological dimension
Extensional causal relation on states
$\boldsymbol{M}$ (γ_{φ}, $\gamma_{\varphi,\psi}$)
$\boldsymbol{M}_{(E)}$(γ_{φ}, $\gamma_{\varphi,\psi}$)
Extensional causal relation on sets of states
[φ] -> [φ] $\bigcap$ [ψ]
[φ] $->_{(E)}$ [φ] $\bigcap$ [ψ]
Intensional causal relation on fluents/events
φ **causes** ψ
E **causes** ψ **if** φ
Facts
Events
Ontological dimension

Figure 8 The multi-faceted causation space.

view, causal laws are primary, and causal relations are secondary. This view can be expressed as *the thesis of the Humean supervenience of causal relations* [42, p. 29]:

> The truth values of all singular causal statements are logically determined by the truth values of statements of causal laws, together with the truth values of non-causal statements about particulars.

If this view is correct, then the fundamental concept is that of a causal law. Consider two possible worlds that agree on all causal laws, and on all non-causal properties of, and relations between, particular states of affairs or events. Then, according to the reductionist position, these two worlds must also agree on all causal relations between states of affairs or events. The opposite singularist viewpoint suggests that causal relations between states of affairs are primary, and causal laws are secondary. C.J. Ducasse [4, p. 129], for example, persuasively argued that

> The causal relation is essentially a relation between concrete individual events; and it is only so far as these events exhibit likeness to others, and can therefore be grouped with them into kinds, that it is possible to pass from individual causal facts to causal laws.

According to an extreme singularist view, it is possible for two events or states of affairs to be causally related without that relation being an instance of any law. This position allowed C. J. Ducasse to advocate, in particular, that causal connection is not an objective connection, and "is not a sensation at all, but a relation ... which has individual concrete events for its terms" [4, p. 132]. We believe, however, that the nomological aspect of causation is orthogonal to the ontological–epistemological plane, and generates another dimension in the multi-faceted causation space (figure 8). This thesis is indirectly supported by the reconciliation of the seemingly polar reductionist and singularist views, suggested by Davidson [3, p. 85], who made a distinction between "knowing there is a law 'covering' two events and knowing what the law is: in my view, Ducasse is right that singular causal statements entail no law; Hume is right that they entail there is a law." In other words, one may define different maps (surjections and/or injections) between causal laws and causal relations in the multi-faceted causation space, and some of these maps may not be identifiable from a single point on a given plane.

The multi-faceted causation space allows us to better position our semantical framework, while avoiding commitment to a particular philosophical stand-point. More precisely, whenever we translate between the causal relation $\mathcal{M}$ and individual causal rules or some relation(s) defined in terms of fluent-triggered causality, we confine the translation(s) to the epistemological dimension in the ontological–epistemological plane. Such translations are not intended to reflect on or discover new causal laws, but rather to re-shape causal relations used by the agent. This is one of the underlying reasons allowing us to entertain a possibility that our motivating approaches can be represented in a unifying setting: they seem to belong to the same surface in the multi-dimensional causation space.

4 An augmented preferential semantics

The introduction of the causal relation $\mathcal{M}$ defined on the information state-space completes our preliminary framework. The only remaining piece is a refinement of the selection function. Our motivation is to retain the preferential style of the semantics considered earlier, and make use of the causal relation in the selection function. Intuitively, we can trace an agent's reasoning about an action e producing successor states $Res(w, e)$ as follows. First of all, some bounded start area (let us call it the *trigger*) is determined in the information state-space Γ – the agent entertains the states in the trigger area as the nearest possible information states compatible with the action's direct effects. Then, the agent begins a propagation from the trigger, driven by the causal relation $\mathcal{M}$. This propagation may explore the whole state-space Γ, but is expected to reach at least one stable state (an element of $\mathcal{K}_{\mathcal{M}}$) – otherwise the action e is qualified, and $Res(w, e) = \emptyset$. The length or configuration of the explored causal chains should not matter, as the propagation process occurs in the information state-space – in other words, a real-time aspect of knowledge dynamics is secondary to our objectives. What is important, however, is that final state(s) are "stable" and "reachable" from the trigger area. If a projection from such a final state to the state-space $\mathcal{W}$ results in a legitimate state r compatible with e (in other words, $r \in \mathcal{D} \cap [e]$), then the state r is a successor state.

Let $\mathcal{M}^*$ be the transitive closure of the relation $\mathcal{M}$. We shall say that an information state β is $\mathcal{M}$-reachable from an information state α, if and only if $\mathcal{M}^*(\alpha, \beta)$. By $\alpha(w)$ we again denote any information state such that $\mathcal{P}(\alpha(w)) = w$. All information states $\alpha(w)$ which project onto the same world state are assigned the same ordering $<_{\alpha(w)}$. Also, let us refer to the set $min(<_{\alpha(w)}, [e]^\Gamma)$ as the trigger area – intuitively, it contains all $<_{\alpha(w)}$-minimal information states whose projections are compatible with the post-conditions of action e. We shall say that a legitimate state r is a successor state, $r \in Res(w, e)$, if and only if r is compatible with the action's direct effects and is a projection of some stable information state, which is $\mathcal{M}$-reachable from a minimal state in the trigger (figure 9). More precisely,

$$\begin{gathered} Res(w, e) = \mathcal{D} \cap [e] \cap \\ \mathcal{X}(\{\rho \in \mathcal{K}_{\mathcal{M}} : \ \mathcal{M}^*(\varphi, \rho), \text{ where } \varphi \in min(<_{\alpha(w)}, [e]^\Gamma)\}). \end{gathered}$$

Given our density condition

$$\mathcal{D} \cap \mathcal{X}(\Gamma \setminus \mathcal{K}_{\mathcal{M}}) = \emptyset,$$

we could omit the requirement $\rho \in \mathcal{K}_{\mathcal{M}}$ in the definition of $Res(w, e)$ – because non-stable states would not be projected onto legitimate states in $\mathcal{D}$ anyway. But it is more intuitive to consider only stable $\mathcal{M}$-reachable states as final points of causal propagation.

This definition clearly separates aspects of minimality and causality in our semantics – the former is captured by a preferential structure $\mathcal{O}$ and the latter by a binary relation $\mathcal{M}$. We claim therefore, that both minimal change and causation are essential in furnishing an efficient solution to the frame and ramification problems. In summary, the *augmented preferential semantics*, can be presented by a tuple $\langle \mathcal{W}, \mathcal{D}, \Gamma, \mathcal{E}, \mathcal{O}, \mathcal{M} \rangle$, where

- $\mathcal{W}$ is the set of world states;
- $\mathcal{D}$ is the set of legitimate world states;
- Γ is the set of information states;
- $\mathcal{E}$ is the set of actions;

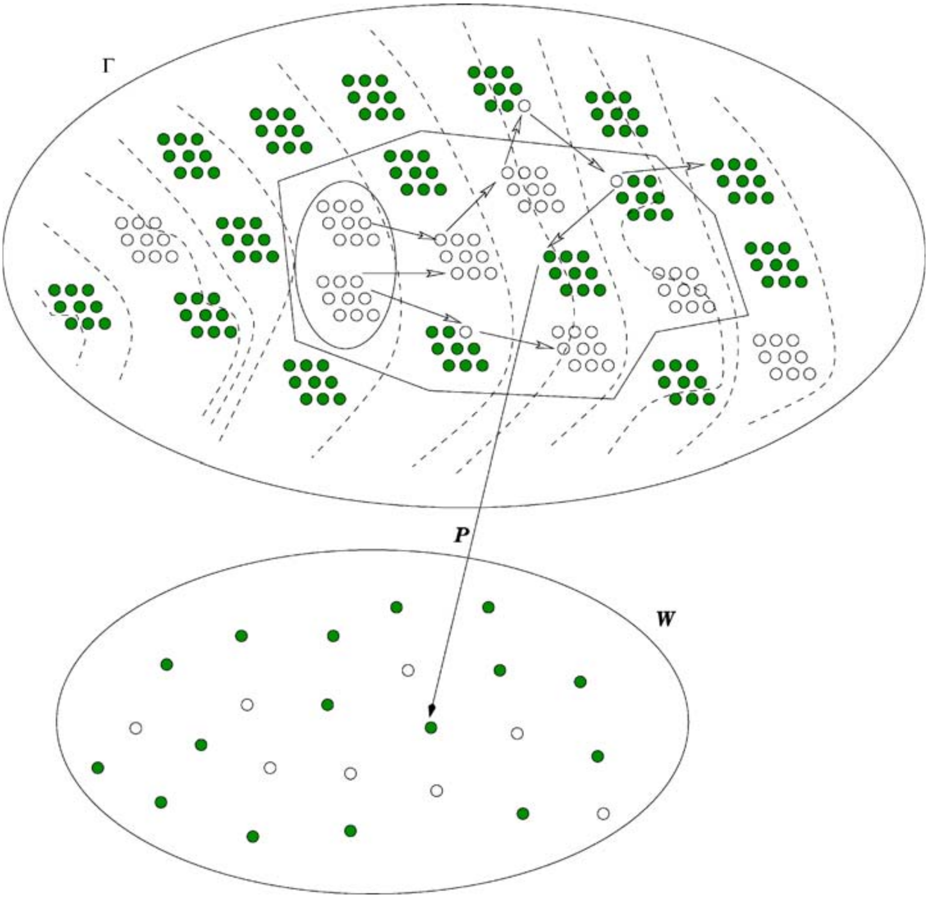


Figure 9 Expanding minimality layers are shown with dashed curves. The polygon encloses the states in $[e]^{\Gamma}$. The trigger area is located within the ellipse, starting causal propagation in the information state-space (*white arrows*), followed by projection onto normal space. The density condition ensures that there are no filled circles with outgoing causal links.

- $\mathcal{O}$ is the preferential structure on $\Gamma \times \Gamma$ (the set of orderings $<_\alpha$ defined with respect to each information state $\alpha \in \Gamma$);
- $\mathcal{M}$ is the causal binary relation on $\Gamma \times \Gamma$,

together with the functions

- Post-condition $[e] : \mathcal{E} \to 2^{\mathcal{W}}$;
- Projection $\mathcal{P}(\gamma) : \Gamma \to \mathcal{W}$;
- Selection $Res(w, e) : \mathcal{W} \times \mathcal{E} \to 2^{\mathcal{W}}$.

Unidirectionality of causation is reflected in the asymmetry condition ($\mathcal{M}_2$): if the state of affairs α causes the state of affairs β, then it cannot be the case that β causes α. This property is closely related to *causal priority* [42] – causes are prior to their effects and, if one presupposes a temporal ordering, causes cannot succeed their effects in time. However, causal priority is not necessarily captured by causal asymmetry as, for example, pointed out by Tooley [42, p. 179]. Tooley considers the ancestral of causal relation that is necessarily transitive (in addition to being asymmetric), and a strict partial ordering defined by this relation on states of affairs:

> But if R is any asymmetric and transitive relation, then the inverse relation, R^*, is also asymmetric and transitive, and it generates an ordering that is indistinguishable from that generated by R, except that it is opposite in direction. This means that any satisfactory account of causal priority, in addition to explaining the asymmetry of causal relations, must also supply some account of why it is the *direction* defined by those relations, rather than that defined by the inverse relations, that is *the* direction of causal processes.

Our motivating approaches do not explicitly use a temporal ordering, and we would like to capture the property of causal priority without relying on some notion of temporal priority. Moreover, it has been argued that causal concepts are more basic than temporal ones, and have tried to analyse the latter in terms of the former. For instance, Mellor [24] posed the question: “The cause–effect relation has no preferred spatial dimension. Why then does it have a temporal one?”, and argued that “causation is what distinguishes time from space and gives it its direction: in short, that time is the causal dimension of spacetime.”

Fortunately, our framework already includes a condition that identifies the causal priority. This surprisingly helpful condition is the requirement that

$$\mathcal{D} \cap \mathcal{X}(\Gamma \setminus \mathcal{K}_{\mathcal{M}}) = \emptyset,$$

Let us show that this density condition affects causal priority and prevents the reversal described by Tooley [42].

First of all, we introduce a *trivially* stable information state – a state which is not related by means of the relation $\mathcal{M}$ with any other information state (neither as cause nor as effect). There are domains where all legitimate states are projections of only trivially stable information states, although the relation $\mathcal{M}$ does not have to be empty. In other words, stable states β that appear as effects in pairs $\mathcal{M}(\alpha, \beta)$ are always projected in this domain onto illegitimate states. We shall refer to such domains as *causally trivial* domains – because causality cannot be used in manipulating legitimate states. On the contrary, in a *causally non-trivial domain*, there is at least one legitimate world state that is a projection of a non trivially stable state.

Now let us consider a causally non-trivial domain. Since the domain is not causally trivial, there is at least one legitimate world state that is a projection of a non trivially stable state. Let $w \in \mathcal{D}$ be such a state. In other words, there must be at least two information states α and β, such that $\mathcal{M}(\alpha, \beta)$, where $\beta \in \mathcal{K}_{\mathcal{M}}$, and $\mathcal{P}(\beta) = w$. Let the relation $\mathcal{M}^{\times}$ be the inverse of $\mathcal{M}$. Hence $\mathcal{M}^{\times}(\beta, \alpha)$, and β is not a stable state with respect to $\mathcal{M}^{\times}$. We obtain that $\beta \notin \mathcal{K}_{\mathcal{M}^{\times}}$, i.e. $\beta \in \Gamma \setminus \mathcal{K}_{\mathcal{M}^{\times}}$. If causal priority is not enforced (i.e., the inverse relation is as causal as the original), then the relation $\mathcal{M}^{\times}$ should also satisfy the requirement

$$\mathcal{D} \cap \mathcal{X}(\Gamma \setminus \mathcal{K}_{\mathcal{M}^{\times}}) = \emptyset.$$

This condition and the observation that $\beta \in \Gamma \setminus \mathcal{K}_{\mathcal{M}^{\times}}$ lead to the conclusion that $\mathcal{P}(\beta) \notin \mathcal{D}$, or in other words, $w = \mathcal{P}(\beta)$ is not a legitimate state. This is a contradiction with our assumption that $w \in \mathcal{D}$. Therefore, the density condition based on the inverse relation $\mathcal{M}^{\times}$ can not hold in a causally non-trivial domain. This fact distinguishes relations $\mathcal{M}$ and $\mathcal{M}^{\times}$, and establishes causal priority. Put simply, causal priority would not matter in causally trivial domains, because then reversals of causal relations would not affect legitimate states and selection functions.

Another related topic is *causal efficacy*: not only are causes prior to their effects but they also "produce" their effects. Following von Wright [44, p. 118]:

> What makes p a cause-factor relative to the effect-factor q is . . . the fact that by *manipulating* p, i.e., by producing changes in it 'at will' as we say, we could bring about changes in q. This applies both to cause-factors which are sufficient and those which are necessary conditions of the corresponding effect-factor.

Causal efficacy and other properties of causation, such as irreflexivity, unidirectionality (asymmetry), and priority, provide grounds for causal propagation along the relation $\mathcal{M}$. In other words, the intuition behind $\mathcal{M}(\alpha, \beta)$ and $\mathcal{M}(\beta, \gamma)$ is that by triggering the information state α, we could bring about changes reflected in the information state β, and that, in turn, produces changes represented in γ, and so on.

An appealing analysis of necessity and sufficiency of causes with respect to their effects was given by Mackie [17]. Instead of defining causes as necessary and/or sufficient conditions, Mackie proposes that, in the relation 'c causes e,' the cause c can be typically represented as "an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result" [17] – the INUS condition. More precisely, 'c causes e' whenever $(c \wedge x) \vee y$ is a necessary and sufficient condition for e, for some x and y, where neither c, x or y is redundant. A frequently cited example, suggested by Mackie in the original work, is a short circuit said to have caused a fire, in the presence of inflammable material, the absence of suitably placed sprinkler, and other relevant conditions. These unnecessary conditions x are sufficient when combined with the short circuit c, which by itself is insufficient. The fire can start in other ways y as well (eg., "overturning of a lighted oil stove") – that is why $(c \wedge x) \vee y$ becomes necessary and sufficient. Shoham [38] supported the intuitive distinction between causes c and other relevant enabling conditions x, but criticised the INUS condition as being too weak formally. The main difference between Mackie's account and Shoham's well-known formalisation, *the logic of chronological ignorance*, was described by Shoham as follows:

> . . . through the use of modal logic, I made those other conditions secondary: whereas causes need to be known (i.e. they are $\Box$-conditions), it is sufficient that the "enabling" conditions not be known to be false (i.e., they are $\Diamond$-conditions).

Causal rules $\Box c \wedge \Diamond x \supset \Box e$, employed by the logic of chronological ignorance, or INUS conditions in general, presuppose causal propagation. Mackie mentioned "some line or chain of causation, some continuous causal process" [16], while the logic of chronological ignorance used "an effective procedure . . . starting with knowledge of only tautologies prior to the earliest point . . . and then iteratively progressing in time, adding only knowledge . . . that is necessitated by prior knowledge and ignorance," while skipping irrelevant points [37].

In summary, causal reasoning is a context-sensitive process, propagating some information triggered by actions and/or events. In the following sections we shall consider action theories dealing with causal propagation in various ways, and then attempt to capture them within our semantics.

5 Thielscher's approach of causal relationships

In this section we review Thielscher's [40] causal relationships approach. Let $\mathcal{F}$ be a finite set of symbols from a fixed language $\mathcal{B}$, called fluent names. A *fluent literal* is either a fluent name $f \in \mathcal{F}$ or its negation, denoted by $\neg f$. Let $L_{\mathcal{F}}$ be the set of all fluent literals defined over the set of fluent names $\mathcal{F}$. A maximal consistent set of fluent literals is called a *state*. For convenience, we denote the set of all states as $\mathcal{W}$ (identifying it with the set of world states in our framework). We shall call the number m of fluent names in $\mathcal{F}$ the dimension of $\mathcal{W}$. By $[\phi]$ we denote all states consistent with the sentence $\phi \in \mathcal{B}$ (i.e., $[\phi] = \{w \in \mathcal{W} : w \vdash \phi\}$). Obviously, our intention is to identify $[\phi]$ with the post-conditions of the action specifying ϕ as its direct effects. Domain constraints are sentences which have to be satisfied in all states. We shall also adopt from Thielscher [40] the following notation. If $\epsilon \in L_{\mathcal{F}}$, then $|\epsilon|$ denotes its *affirmative component*, that is, $|f| = |\neg f| = f$, where $f \in \mathcal{F}$. This notation can be extended to sets of fluent literals as follows: $|S| = \{|f| : f \in S\}$.

Thielscher's [40] causal theory of action consists of two main components: *action laws* which describe the direct effects of actions performed in a given state, and *causal relationships* which determine the indirect effects of actions. Every action law[4] contains a condition C, which is a set of fluent literals, all of which must be contained in an initial state where the action is intended to be applied; and a (direct) effect E, which is also a set of fluent literals, all of which must hold in the resulting state after having applied the action. Condition and effect are constructed from the same set of fluent names so that the state obtained from a direct effect is determined by removing C from the initial state and adding E to the result. An action may result in a number of state transitions.

Definition 5.1 Let $\mathcal{F}$ be the set of fluent names and let $\mathcal{A}$ be a finite set of symbols called action names, such that $\mathcal{F} \cap \mathcal{A} = \emptyset$. An action law is a triple $\langle C, a, E\rangle$ where C, called *condition*, and E, called *effect*, are individually consistent sets of fluent literals, composed of the very same set of fluent names (i.e., $|C| = |E|$) and $a \in \mathcal{A}$. If w is a state, then an action law $\alpha = \langle C, a, E\rangle$ is applicable in w if and only if $C \subseteq w$. The application of α to w yields the state $(w \setminus C) \cup E$ (where $\setminus$ denotes set subtraction).

Causal relationships are specified as ϵ *causes* ρ *if* Φ, where ϵ and ρ are fluent literals and Φ is a fluent formula based on the set of fluent names $\mathcal{F}$. Causal relationships are derived from a set of domain constraints D, using predefined *influence* dependencies between certain fluents. Another important concept is a state-effect pair (s, E) containing a state s and a set of fluent literals E. The second component is used to "record" a (partial) history of fluents that changed their values in transitions leading to the state represented by the first component of the pair – more precisely, the "history" contains only the latest (current) values of all the changed fluents, making it a snapshot of current effects.

Definition 5.2 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) if and only if $\Phi \wedge \neg\rho$

[4] Again, action laws should be more appropriately called action rules.

is true in s, and $\epsilon \in E$. Its application yields the pair (s', E'), denoted as $(s, E) \leadsto (s', E')$, where $s' = (s \setminus \{\neg\rho\}) \cup \{\rho\}$ and $E' = (E \setminus \{\neg\rho\}) \cup \{\rho\}$.

In other words, a causal relationship is applicable if Φ holds at the current state s, the indirect effect ρ is false and the cause ϵ is among the current effects E. Note that ϵ must be among the current effects; being true at the current state is not sufficient.

A possible *successor state* is determined through repeated application of causal relationships. This process may pass through states violating domain constraints. This is permissible only if subsequent applications of causal laws result in legal states. Specifically, given an initial state w and action a, the set of possible successor states $Res_{RD\mathcal{L}}(w, a)$ is determined as follows.

Definition 5.3 ($Res_{R,D,\mathcal{L}}(w, a)$) Let $\mathcal{F}$ be the set of fluent names, A a set of action names, $\mathcal{L}$ a set of action laws, D a set of domain constraints, and R a set of causal relationships. Furthermore, let w be a state satisfying D and let $a \in A$ be an action name. A state r is a *successor state* of w and a, denoted $r \in Res_{RD\mathcal{L}}(w, a)$, if and only if there exists an applicable (with respect to w) action law $\alpha = \langle C, a, E\rangle \in \mathcal{L}$ such that

1. $((w \setminus C) \cup E, E) \overset{*}{\leadsto} (r, E')$ for some E', and
2. r satisfies D,

where $\overset{*}{\leadsto}$ denotes the transitive closure of $\leadsto$.

Since R is derived from D, a successor state satisfying domain constraints cannot belong to a pair where a causal relationship is applicable. We can strengthen the concept of successor states to *conservative successor states*, denoted $Res^*_{R,D,\mathcal{L}}(w, a)$, which are resultant states satisfying the direct effects of the action, as follows.

Definition 5.4 ($Res^*_{R,D,\mathcal{L}}(w, a)$) Let $\mathcal{F}$, A, $\mathcal{L}$, D, R, w, $\alpha = \langle C, a, E\rangle$ be as in Definition 5.3. A state r is a *conservative successor state* of w and a, $r \in Res^*_{R,D,\mathcal{L}}(w, a)$, if and only if

1. $r \in Res_{R,D,\mathcal{L}}(w, a)$, and
2. $E \subseteq r$.

Using this definition, causal propagation can "travel" outside E-states, however it must end in a state consistent with the direct effects E. As mentioned before, the occurrence of a literal ϵ in a state s does not guarantee that a causal relationship ϵ *causes* ρ *if* Φ is applicable to a pair (s, E) – to ensure applicability, the literal ϵ has to belong to the current effects E. That is why, in order to trace causal propagation with causal relationships, one needs to keep an explicit (and changing) account of context-dependent effects of actions. Interestingly, given a *transition* state-effect pair (s, E), if the literal ϵ is part of the current effects E, then it must be an element of the current state s. This observation can be formalised as follows.

Lemma 5.5 If $E \subseteq s$ and $(s, E) \overset{*}{\leadsto} (s', E')$, then $E' \subseteq s'$.

6 McCain and Turner's causal fixed-points approach

In this section we sketch McCain and Turner's [18] causal theory of actions, while reusing some notation from the previous section. McCain and Turner introduce a new connective $\Rightarrow$ to denote a causal relationship between sentences ϕ and ψ of the underlying language $\mathcal{B}$. This allows for expressions of the form $\phi \Rightarrow \psi$ (where $\phi,\ \psi \in \mathcal{B}$) which are termed *causal laws* (or *causal rules* – we prefer the latter term, for the reasons mentioned earlier). Nesting of $\Rightarrow$ is not permitted, and the antecedent of any causal rule is assumed to be consistent. A set of causal rules $\mathcal{Q}$ is referred to as a *causal system*. Given any set of sentences $\Lambda \subseteq \mathcal{B}$ and a causal system $\mathcal{Q}$, the causal *closure* of Λ in $\mathcal{Q}$ is denoted $C_{\mathcal{Q}}(\Lambda)$ and defined to be the smallest superset of Λ closed under classical logical consequence such that for any $\phi \Rightarrow \psi \in \mathcal{Q}$, if $\phi \in C_{\mathcal{Q}}(\Lambda)$, then $\psi \in C_{\mathcal{Q}}(\Lambda)$. We also say that Λ *causally implies* ϕ with respect to $\mathcal{Q}$ if and only if $\phi \in C_{\mathcal{Q}}(\Lambda)$ and denote this as $\Lambda \mathrel{|\!\!\sim}_{\mathcal{Q}} \phi$. Any state r is legitimate with respect to $\mathcal{Q}$ if and only if $r = C_{\mathcal{Q}}(r) \cap L_{\mathcal{F}}$. That is, a state is legitimate if and only if it does not contravene any causal laws of $\mathcal{Q}$.

McCain and Turner's aim is to determine the set of possible resultant (or successor) states $Res_{\mathcal{Q}}(w, e)$ given an initial state w and the direct effects (or postconditions) of an action represented by the sentence e. Formally, for any causal system $\mathcal{Q}$, we have a function $Res_{\mathcal{Q}}$ mapping a legitimate initial state w and sentence e (direct effects) to the set of states $Res_{\mathcal{Q}}(w, e)$ according to the definition [18]:

$$r \in Res_{\mathcal{Q}}(w, e) \quad \text{if and only if} \quad r = \{p \in L_{\mathcal{F}} :\ (w \cap r) \cup \{e\} \mathrel{|\!\!\sim}_{\mathcal{Q}} p\}$$

The elements of $Res_{\mathcal{Q}}(w, e)$ are referred to as *causal fixed-points*. Intuitively, the elements of $Res_{\mathcal{Q}}(w, e)$ are simply those e-states where all changes with respect to w can be justified by the underlying causal system. In short, every literal in the successor state must be "explained" either as persisting through the action, or as a direct effect, or as a causal ramification of other literals of the successor state. It was emphasised previously that this definition captures the causal minimisation policy: the world changes as little as *necessary* when an action is performed.

Importantly, every causal fixed-point is a successor according to the causal relationship approach (but not vice versa), using a simple construction of causal relationships from causal rules [40]. Therefore, if the augmented preferential semantics captures causal relationships, it would capture causal fixed-points as well.

7 Representation results

We maintain that causal propagation is employed, as a process exploring the information state-space in a search for necessary and/or sufficient conditions, in both these approaches. The differences, we believe, are due to different ways taken to handle the *context-sensitive* nature of causal propagation. In particular, the causal relationships of Thielscher create causal histories that affect simple propagation, because the process must keep an account of all changes. The causal theories of McCain and Turner "warp" simple propagation by restricting successor states to causal fixed-points. Arguably, a better unifying alternative is to encode causal context in the preferential structure and causal relation, present in our semantics.

7.1 Propagating minimal change with causal relationships

Our intention at this stage is to consider a formalisation which faithfully captures conservative successor states, defined by $Res^{*}_{R,D,\mathcal{L}}(w, a)$, using a selection mechanism based on the augmented preferential semantics. More precisely, we would like to use a binary (causal) relation on states, while propagating within a set of possible states, instead of keeping an explicit (and changing) account of context-dependent effects of actions. The advantage of this proposal is that a causal relation would be action-independent, unlike the history of effects. Obviously, this objective is hardly achievable without extending the action systems components in some way – in particular, we cannot use the approximation $\mathcal{W} = \Gamma$. In other words, we make use of the information state-space Γ explicitly. The construction of this space is inspired by the two-steps construction – first, via the "hyper-state space," following the techniques reported by Prokopenko et al. [28], and then via the "power-state space," introduced by Prokopenko et al. [29]. However, the new process presented here is much simpler.

We begin by adding to the set of fluent names $\mathcal{F}$, and constructing the set of *justifier fluents* $\mathring{\mathcal{F}}$ which has the same cardinality as $\mathcal{F}$ (i.e., for each fluent $f \in \mathcal{F}$ we add an additional fluent $\mathring{f}$). The justifier fluents are intended to maintain contextual information that becomes important during causal propagation. As indicated earlier, $L_{\mathcal{F}}$ is the set of all fluent literals defined over the set of fluent names $\mathcal{F}$. The set of *justifier literals* $\mathring{L}_{\mathcal{F}} = \mathring{\mathcal{F}} \cup \{\neg q : q \in \mathring{\mathcal{F}}\}$ is mapped from $L_{\mathcal{F}}$ by the function $l : L_{\mathcal{F}} \to \mathring{L}_{\mathcal{F}}$. We will use the abbreviation $\mathring{f}$ instead of $l(f)$ for simplicity. In addition, a justifier set $\mathring{J}$ for any set of fluent literals $J \subseteq L_{\mathcal{F}}$ is defined as $\mathring{J} = \cup_{f \in J}\{\mathring{f}\}$. These definitions allow us to construct an information state.

Definition 7.1 (Information state) Given a set of fluents $\mathcal{F}$, an *information state* is a union of a maximal consistent set of literals from $L_{\mathcal{F}}$ and a consistent set of literals from $\mathring{L}_{\mathcal{F}}$.

Note that justifier literals are not required to form a *maximal* consistent set. For example, given the set of fluents $\mathcal{F} = \{a, b\}$, one may construct nine information states for each maximal consistent set of literals from $L_{\mathcal{F}}$, i.e., for each state in $\mathcal{W}$. In particular, for the maximal consistent set of literals $\{a, b\}$, these information states are possible:

$$\{a, b\},\ \{a, b, \mathring{a}\},\ \{a, b, \mathring{b}\},\ \{a, b, \neg\mathring{a}\},\ \{a, b, \neg\mathring{b}\},$$
$$\{a, b, \mathring{a}, \mathring{b}\},\ \{a, b, \neg\mathring{a}, \mathring{b}\},\ \{a, b, \mathring{a}, \neg\mathring{b}\},\ \{a, b, \neg\mathring{a}, \neg\mathring{b}\}$$

A justifier fluent $\mathring{f}$ is able to take *three* values: "positive" $\mathring{f}$, "negative" $\mathring{f}$ and "unknown" – the latter form is indicated by the absence of both $\mathring{f}$ and $\neg\mathring{f}$ in the information state. Intuitively, presence of either justifier literal $\mathring{f}$ or $\neg\mathring{f}$ in an information state points to a known cause (positive or negative) for the literal f, while absence of a justifier literal indicates that a cause is unknown. Given these three possibilities for justifier literals, for each maximal consistent set of literals from $L_{\mathcal{F}}$, there are 3^m information states, where m is the dimension of the set of fluents $\mathcal{F}$. There are 2^m states in $\mathcal{W}$, and therefore, the information state-space Γ constructed in this way has the cardinality of $2^m 3^m$. The following two definitions map the information-state space Γ to the normal space $\mathcal{W}$ and vice versa.

Definition 7.2 (Projection from information-state space) A projection from Γ to $\mathcal{W}$, $\mathcal{P} : \Gamma \to \mathcal{W}$, is the function mapping an information state $s \in \Gamma$ to a state $r \in \mathcal{W}$ as follows: $r = s \cap L_{\mathcal{F}}$.

Each information state $s \in \Gamma$ has also a (possibly empty) justifier subset, defined as $j(s) = s \setminus \mathcal{P}(s)$, or simply $j(s) = s \cap \overset{\circ}{L}_{\mathcal{F}}$. If $j(s) = \emptyset$, or in other words, the information state s and its projection $r = \mathcal{P}(s)$ contain exactly the same literals, we shall call the information state s the *mirror* of the state r, and denote it as $\mu(r)$. Formally, the function $\mu : \mathcal{W} \to \Gamma$ maps a state r to an information state s if and only if $r = \mathcal{P}(s)$ and $j(s) = \emptyset$. Of course, *in terms of the literals* contained in r and its information mirror s, these two states are simply identical – but they belong to two different spaces $\mathcal{W}$ and Γ.

Definition 7.3 (Information-neighbourhood) An *information-neighbourhood* of a state $r \in \mathcal{W}$ is the function $N : \mathcal{W} \to \Gamma$, mapping a state r to a set of information states: $N(r) = \{s \in \Gamma : r = \mathcal{P}(s)\}$.

Intuitively, the set $N(r)$ contains information states where all possible causes (i.e., justifier literals) vary, while the (proper) literals defined on $\mathcal{F}$ are fixed. By definition, the mirror state $\mu(r) \in N(r)$. The combinatorial variability of possible causes in an information-neighbourhood allows us to account for different action-dependent histories in a causally propagated chain, leading to a successor state in $Res^*_{R,D,\mathcal{L}}(w, a)$. Before we formally specify a binary causal relation on information states, let us illustrate its purpose. Suppose we have an action system with $\mathcal{F} = \{a, b\}$, $D = \{\neg b \to \neg a\}$, $R = \{\neg b \ \textit{causes} \ \neg a \ \textit{if} \ \top\}$, and $\mathcal{L} = \{\langle\{b\}, x, \{\neg b\}\rangle\}$. Let us consider the action x executed in the initial state $w = \{a, b\}$. The action's direct effect is $\{\neg b\}$, yielding the intermediate state $\{a, \neg b\} = (w \setminus \{b\}) \cup \{\neg b\}$. This state contradicts the given domain constraint. However, the system's sole causal law applies: $(\{a, \neg b\}, \{\neg b\}) \leadsto (\{\neg a, \neg b\}, \{\neg a, \neg b\})$. The state component of the resultant pair obeys the domain constraint (and satisfies the direct effect, in addition). Therefore, it belongs to $Res^*_{R,D,\mathcal{L}}(w, x)$ – a singleton set.

We now indicate how this propagation can be traced in the information state space. The information-neighbourhood $N(q)$ of the intermediate state $q = \{a, \neg b\}$ contains the information state which explicitly represents the initial history component $\{\neg b\}$ – this information state is exactly $\{a, \neg b, \neg \overset{\circ}{b}\}$. The information-neighbourhood of the successor state $q' = \{\neg a, \neg b\}$ contains the state accountable for the final history component $\{\neg a, \neg b\}$. This state is exactly $\{\neg a, \neg b, \neg \overset{\circ}{a}, \neg \overset{\circ}{b}\}$. The idea, then, is to construct just such a binary relation on information states so that transitions in the information state space correspond to causal propagation.

Definition 7.4 (Binary relation $\rightharpoonup$) A binary relation $\rightharpoonup$ is defined on $\Gamma \times \Gamma$. Given two elements $x_1, x_2 \in \Gamma$, we say that $x_1 \rightharpoonup x_2$ if and only if there exists a causal relationship ϵ *causes* ρ *if* Φ such that

1. $\mathcal{P}(x_1) \vdash \epsilon \wedge \Phi \wedge \neg\rho$
2. $j(x_1) \vdash \overset{\circ}{\epsilon}$
3. $\mathcal{P}(x_2) = (\mathcal{P}(x_1) \setminus \{\neg\rho\}) \cup \{\rho\}$
4. $j(x_2) = (j(x_1) \setminus \{\neg\overset{\circ}{\rho}\}) \cup \{\overset{\circ}{\rho}\}$

That is,

1. The causal relationship is applicable to all pairs $(\mathcal{P}(x_1), E)$, where $\epsilon \in E$;
2. The antecedent ϵ of the causal relationship is among the current effects – and therefore, the justifier literal $\overset{\circ}{\epsilon}$ belongs to the information state x_1, i.e., $\overset{\circ}{\epsilon} \in x_1$;
3. The state $\mathcal{P}(x_2)$ is like $\mathcal{P}(x_1)$ but ρ has changed value;
4. The effect ρ is added to the current effects – and therefore, the justifier $\overset{\circ}{\rho}$ is included in the information state x_2.

A single causal relationship may generate a number of transitions. For example, the causal relationship $\neg b$ *causes* $\neg a$ *if* $\top$ generates, among others, the following three transitions between information states in the neighbourhoods of the states $q = \{a, \neg b\}$ and $q' = \{\neg a, \neg b\}$:

$$\{a, \neg b, \neg\overset{\circ}{b}\} \rightharpoonup \{\neg a, \neg b, \neg\overset{\circ}{a}, \neg\overset{\circ}{b}\}$$

$$\{a, \neg b, \overset{\circ}{a}, \neg\overset{\circ}{b}\} \rightharpoonup \{\neg a, \neg b, \neg\overset{\circ}{a}, \neg\overset{\circ}{b}\}$$

$$\{a, \neg b, \neg\overset{\circ}{a}, \neg\overset{\circ}{b}\} \rightharpoonup \{\neg a, \neg b, \neg\overset{\circ}{a}, \neg\overset{\circ}{b}\}$$

7.2 Representation results for causal relationships

We intend to show that it is possible to correctly capture successor states obtained by the causal relationship approach, using causal propagation in the information-state space Γ driven by the binary relation $\rightharpoonup$. Informally, this process simply propagates "minimal change," instead of keeping an explicit and changing account of context-dependent effects of actions. Let $\overset{*}{\rightharpoonup}$ denote the transitive closure of the binary relation $\rightharpoonup$. The following lemma characterises chains of causal relationships in terms of transitions $x \rightharpoonup x'$.

Lemma 7.5 For $x, y \in \Gamma$ and a set of fluent literals E such that $E \subseteq \mathcal{P}(x)$, $x \overset{*}{\rightharpoonup} y$ if and only if $(\mathcal{P}(x), E) \overset{*}{\rightsquigarrow} (\mathcal{P}(y), T)$, where $\overset{\circ}{E} = j(x)$ and $\overset{\circ}{T} = j(y)$.

We introduce here one more useful definition. A *trigger* information state is the state s where the projection $p(s)$ identifies the nearest states (to the initial state w), among states consistent with the direct effects E, and justifier literals in $j(s)$ capture the immediate causal context.

Definition 7.6 (Trigger information state) An information state $\|E\|_w \in \Gamma$ is a trigger state for an initial state $w \in \mathcal{W}$ and an action a with the action law $\langle C, a, E\rangle$, if and only if

$$\mathcal{P}(\|E\|_w) \in min(\prec_w, [E]) \text{ and } j(\|E\|_w) = \overset{\circ}{E},$$

where $\prec_w$ is the PMA ordering.

That is, in terms of the PMA ordering, the projection of $\|E\|_w$ is nearest to the initial state w among the post-condition states $[E]$. Since E is a set of literals, it is easy to see that $(\mathcal{P}(\|E\|_w) = (w \setminus C) \cup E$. Importantly, the trigger state represents the initial (immediate) causal context, i.e., initial causally justified changes triggered by effects E. For instance, if the action law $\langle\{b\}, x, \{\neg b\}\rangle$ is applied to the initial state $\{a, b\}$, then the trigger state $\|\{\neg b\}\|_{\{a,b\}}$ is given by $\{a, \neg b, \neg\overset{\circ}{b}\}$. In the information-state space causal propagation starts from the trigger information state, or, in other words, from the state reflecting only immediate causes $(\neg b)$ and their justifier literals $(\neg\overset{\circ}{b})$. We can view changes triggered by the state $\|E\|_w$ as propagating "minimal change" in information-state space towards a stable information state that can be projected onto a possible successor state.

The last step is a construction of the preferential structure $\mathcal{O}$. The intention is to construct orderings $\ll$ in $\mathcal{O}$ in such a way that, given an initial state w and an action with the post-condition E, the trigger information state $\|E\|_w$ is a $\ll$-minimal state among all information states in $[E]^\Gamma$. The following lemma ensures this possibility.

Lemma 7.7 For a state $w \in \mathcal{W}$, there exists an ordering $\ll$ with respect to the mirror information state $\mu(w)$, such that for any action law $\langle C, a, E\rangle$,

$$\{\|E\|_w\} = min(\ll_{\mu(w)}, [E]^\Gamma).$$

Lemmas 7.5 and 7.7 contribute to the following main result.

Theorem 7.8 *For every action system based on causal relationships there exists a selection-equivalent action system* $\langle\mathcal{W}, \mathcal{D}, \Gamma, \mathcal{E}, \mathcal{O}, \mathcal{M}\rangle$.

Conversely, for every action system $\langle\mathcal{W}, \mathcal{D}, \Gamma, \mathcal{E}, \mathcal{O}, \mathcal{M}\rangle$ *there exists a selection-equivalent action system based on causal relationships.*

In establishing the desired selection-equivalence we use the space Γ, specify $\mathcal{M} = \rightharpoonup$, and include in the set $\mathcal{E}$ all the actions a from the set of action laws A, where each action law has the form $\langle C, a, E\rangle$, E being the direct effect of a. The set $\mathcal{D}$ is chosen as a subset of $\mathcal{W}$ such that its elements satisfy the constraints D.

7.3 Further reductions

Since Thielscher's approach handles actions with only conjunctive effects, there is only one intermediate state, nearest to the initial one, from which causal propagation may start. However, the approach reported by Thielscher [41] extends the original one [40] towards *alternative* effect propositions, where the disjunction of effects $e_1 \vee e_2$ is interpreted as exclusive, and inclusive disjunction is simply modelled as $e_1 \vee e_2 \vee (e_1 \wedge e_2)$. In this extended case, alternative effects lead to alternative intermediate (preliminary) states, and the original causal relationship approach is then applied to each of these preliminary states. In other words, in order to account for indirect effects, preliminary states are "taken as starting points for the successive applications of causal relationships" until overall satisfactory successor states are obtained [41]. Obviously, this extension with disjunctive direct effects is also handled by our augmented preferential semantics – because there is no requirement that the set of minimal states, from which the propagation starts, is a singleton.

There also exists an alternative representation for causal fixed-points in McCain and Turner's causal theories of actions [30, 31]: it does not require a higher-dimensional information state-space, but replaces minimal states in the trigger area with a more complex set of *e*-predecessors. The *e*-predecessors of a state r are the states that are closer (in terms of a preference relation, for example, the PMA ordering) to the initial state, where the action e is performed, than r. The employed propagation is more convoluted as well: all the predecessors must be traversed through during the causal propagation. In short, there is a trade-off between dimensionality of the required information state-space and the context-sensitive nature of causal propagation: in a lower-dimensional state-space the context plays a more significant role, forcing "minimal change" to propagate in a non-trivial way.

We also point out that it is possible to represent Sandewall's causal propagation semantics (the *transition cascade semantics*) [35] as an instance of the augmented preferential semantics. While there are certain similarities between the causal propagation semantics and our augmented preferential semantics (including components such as $\mathcal{W}$, $\mathcal{E}$, $\mathcal{D}$, and the causal transition relation), the main difference is that the Principal of Minimal Change is not explicit in the causal propagation semantics. However, we believe that it is hidden behind another concept introduced by Sandewall: an *action invocation* relation $G(e, r, r')$, where e is an action, r is the state where the action e is invoked, and r' is "the new state where the instrumental part of the action has been executed" [35]. A transition cascade, driven by the causal relation, starts from such *invoked* states. An ideal solution would suggest an ordering on states such that the invocation relation G can be simply realised by selecting the nearest states satisfying the post-conditions of the action. However, this does not appear to be possible without restricting the relation G, as demonstrated by Prokopenko et al. [29]. The restrictions ensure that, given an initial state and an action, the invoked states can be characterised precisely as states nearest to the initial one in terms of some appropriate minimality ordering. In other words, the Principle of Minimal Change is put to work instead of the action invocation relation. The restrictions imposed on the invocation relation G use the approximation $\mathcal{W} = \Gamma$ [29, 31] – again, manifesting the trade-off between dimensionality of the required information state-space and the context-sensitive nature of causal propagation.

8 Conclusion and future work

In this paper we examined the role of causality in reasoning about action and change. Our investigation covered a variety of semantics for causal reasoning about action and change – ranging from pure preferential semantics to its variants augmented with a causal relation. In our search we committed neither to a particular philosophic stand-point on the metaphysics of causation or minimality nor to any specific logical language. This approach allowed us to explore the role of several fundamental underlying principles transparently, on a set-theoretic basis, without obscuring these principles by the internal structure of system components. The investigation culminated in a general unifying semantics for a broad class of causal action theories, represented by a number of recent influential approaches.

The reasons behind our expectations that these approaches can be represented in a unifying setting were clarified during construction and analysis of the multi-dimensional causation space (figure 8). Placing our semantical framework in the ontological–epistemological plane of this space allowed us to characterise and contrast our motivating approaches systematically. The unifying semantics describes causal reasoning about action as the propagation of an action's effects from minimal states (determined with respect to a chosen *preferential structure*) to stable states (ascertained with respect to some *causal relation*). Importantly, we identified the concept of the *information state-space*, where the propagation process takes place. This concept accentuates the differences between system dynamics (related to transitions between world states) and knowledge dynamics (involving transitions between information states). In addition, we established a valuable *density* condition linking legitimate system states and stable information states. The density condition ensures proper causal priority of the binary relation used in propagation.

The *augmented preferential semantics*, emerging as a result of this study, captures the causal relationship approach of Thielscher and the causal fixed-points framework of McCain and Turner. It has been also related to the causal propagation semantics of Sandewall in an earlier study, subsuming it under certain uniformity assumptions [29–31]. In summary, this semantics captures two fundamental underlying principles – the *Principle of Minimal Change* and the *Principle of Causal Change* – and illustrates their clear and distinct roles. It is hoped that the unifying semantics would provide further insights into the views on causation and minimality. For example, it explains a distinction between *minimisable* dynamic systems (that can be described by theories of action based on minimal change) and those systems that require *causal theories of action*, or in general, the systems "capable of forms of reasoning that cannot be captured by the Principle of Minimal Change," observed by Pagnucco and Peppas [25]. Their framework included a number of formal properties serving as necessary and sufficient conditions under which a dynamic system is minimisable, establishing the range of applicability of the Principle of Minimal Change. Interestingly, the McCain and Turner theory of causal fixed-points was shown to be minimisable, while the causal relationship approach of Thielscher was not. In particular, the causal fixed-points were characterised purely via a preferential structure, and without causal propagation – however, the employed preferential structure was defined over meta-states of a higher dimension. These meta-states played the role identical to information states in our semantics. In other words, the results presented by Pagnucco and Peppas [25] described an alternative representation of the McCain and Turner theory in line with our semantics – this time, with the empty causal relation $\mathcal{M} = \emptyset$ and $\mathcal{W} \neq \Gamma$. Importantly, the alternative representation [25] demonstrates that the augmented preferential semantics may cover both minimisable and non-minimisable systems. The natural question is, whether this is too broad, and whether one would not be better off concentrating on non-minimisable (causal) systems independently from minimisable ones. While this intention is definitely praiseworthy, we believe that the unifying framework presented in this paper provides an efficient way to compare minimisable and non-minimisable systems. For instance, the systems with causal fixed-points and the systems based on causal relationships definitely differ with respect to being minimisable. In particular, the alternative representation of the former (based on $\mathcal{M} = \emptyset$ and $\mathcal{W} \neq \Gamma$) can be clearly differentiated now from the representation of the latter, that needed both $\mathcal{M} \neq \emptyset$ and $\mathcal{W} \neq \Gamma$. That is, there is a trade-off between these two components. These comments bring us closer to the subject of conciseness. As pointed out by Pagnucco and Peppas [25],

> …if the class of domains at focus is within the range of applicability of both causal and minimal change approaches, the determining factor in choosing between the two could be the "information cost" associated with the usage of each approach.

A unifying semantics for a class of *concise* minimisable and non-minimisable causal logics of action appears at this stage to be one of the most appealing and engaging themes in Reasoning about Action and Change.

Acknowledgements Parts of the results presented here have been taken from the author's doctoral thesis [31]. The author is thankful to his supervisors Pavlos Peppas, Maurice Pagnucco and Abhaya C. Nayak for their generous insights, encouragement and support, and to the anonymous referees for their comments and suggestions that have helped to improve the paper.

Appendix – Proofs

Lemma 5.5 If $E \subseteq s$ and $(s, E) \overset{*}{\leadsto} (s', E')$, then $E' \subseteq s'$.

Proof Assume

$$E \subseteq s, \tag{1}$$

and $(s, E) \leadsto (s', E')$. By Definition 5.2, if $(s, E) \leadsto (s', E')$, then

$$s' = (s \setminus \{\neg\rho\}) \cup \{\rho\} \tag{2}$$

$$E' = (E \setminus \{\neg\rho\}) \cup \{\rho\}. \tag{3}$$

Both propagated components – current state and current effects – are updated with respect to the effect ρ simultaneously and analogously. Using Eqs. 1, 2 and 3, we obtain $E' \subseteq s'$. A simple induction on a length of a $(s_0, E_0) \leadsto (s_1, E_1) \leadsto ... \leadsto (s_n, E_n)$ shows that $E_n \subseteq s_n$. Therefore, the property $E' \subseteq s'$ holds for every pair (s', E') if $(s, E) \overset{*}{\leadsto} (s', E')$ and $E \subseteq s$. □

Lemma 7.5 For $x, y \in \Gamma$ and a set of fluent literals E such that $E \subseteq \mathcal{P}(x)$, $x \overset{*}{\rightharpoondown} y$ if and only if $(\mathcal{P}(x), E) \overset{*}{\leadsto} (\mathcal{P}(y), T)$, where $\overset{\circ}{E} = j(x)$ and $\overset{\circ}{T} = j(y)$.

Proof ($\Longrightarrow$) Let $x \overset{*}{\rightharpoonup} y$ for two states $x, y \in \Gamma$. We need to show that $(\mathcal{P}(x), E) \overset{*}{\rightsquigarrow} (\mathcal{P}(y), T)$, where $\overset{\circ}{E} = j(x)$ and $\overset{\circ}{T} = j(y)$. Since $\overset{*}{\rightharpoonup}$ is a transitive closure of $\rightharpoonup$, there is a sequence of states in Γ, $x_1, ..., x_n$, such that

$$x_1 = x, \;\; x_n = y, \;\; \text{and} \;\; x_i \rightharpoonup x_{i+1} \;\; \text{for} \;\; 1 \leq i < n. \tag{4}$$

According to Definition 5.2, in order to prove the lemma, we need to show that there exists a sequence of causal relationships $cr_1, ..., cr_{n-1}$, such that $cr_i : \epsilon_i$ *causes* ρ_i *if* Φ_i, and cr_i is applicable to the pair (q_i, E_i), yielding (q_{i+1}, E_{i+1}), where $q_i = \mathcal{P}(x_i)$, $\overset{\circ}{E_i} = j(x_i)$, $\{\rho_i\} = q_{i+1} \setminus q_i$, $E_{i+1} = (E_i \setminus \{\neg\rho_i\}) \cup \{\rho_i\}$, and $E_1 = E$, $E_n = T$.

We prove it by induction on the length of this sequence. Let $n = 2$, $q_1 = \mathcal{P}(x_1)$ and $q_2 = \mathcal{P}(x_2)$. Let also the sets of literals E_1 and E_2 be such that $\overset{\circ}{E_1} = j(x_1)$ and $\overset{\circ}{E_2} = j(x_2)$.

Since $x_1 \rightharpoonup x_2$, Definition 7.4 of $\rightharpoonup$ immediately yields that there exists a causal relationship ϵ_1 *causes* ρ_1 *if* Φ_1 such that

1. $q_1 \vdash \epsilon_1 \wedge \Phi_1 \wedge \neg\rho_1$
2. $\overset{\circ}{E_1} \vdash \overset{\circ}{\epsilon_1}$
3. $q_2 = (q_1 \setminus \{\neg\rho_1\}) \cup \{\rho_1\}$
4. $\overset{\circ}{E_2} = (\overset{\circ}{E_1} \setminus \{\neg\overset{\circ}{\rho_1}\}) \cup \{\overset{\circ}{\rho_1}\}$

The second fact means that $\epsilon_1 \in E_1$, and the first two facts together imply that the causal relationship is applicable to (q_1, E_1). The last fact means that $E_2 = (E_1 \setminus \{\neg\rho_1\}) \cup \{\rho_1\}$, and the last two facts together imply that the causal relationship yields (q_2, E_2). In other words, $(q_1, E_1) \rightsquigarrow (q_2, E_2)$ or $(\mathcal{P}(x_1), E_1) \rightsquigarrow (\mathcal{P}(x_2), E_2)$. Trivially, $(\mathcal{P}(x_1), E_1) \overset{*}{\rightsquigarrow} (\mathcal{P}(x_2), E_2)$.

Consider the case of length k. Let us assume that for a sequence of states in Γ, $x_1, ..., x_k$, such that $x_1 \overset{*}{\rightharpoonup} x_k$, there is a sequence of causal relationships $cr_1, ..., cr_{k-1}$, underlying $(q_1, E_1) \overset{*}{\rightsquigarrow} (q_k, E_k)$, where $q_i = \mathcal{P}(x_i)$, for $1 \leq i < k$. Assume also that $x_k \rightharpoonup x_{k+1}$. The argument analogous to the $n = 2$ case shows that there is a causal relationship $cr_k : \epsilon_k$ *causes* ρ_k *if* Φ_k, underlying propagation $(q_k, E_k) \rightsquigarrow (q_{k+1}, E_{k+1})$, where $E_{k+1} = (E_k \setminus \{\neg\rho_k\}) \cup \{\rho_k\}$. By transitivity of $\rightsquigarrow$, we obtain immediately $(q_1, E_1) \overset{*}{\rightsquigarrow} (q_{k+1}, E_{k+1})$ or $(\mathcal{P}(x_1), E_1) \overset{*}{\rightsquigarrow} (\mathcal{P}(x_2), E_{k+1})$.

This proves that $(\mathcal{P}(x), E) \overset{*}{\rightsquigarrow} (\mathcal{P}(y), T)$, where $\overset{\circ}{E} = j(x)$ and $\overset{\circ}{T} = j(y)$.

($\Longleftarrow$) Let $E \subseteq \mathcal{P}(x)$ and $(\mathcal{P}(x), E) \overset{*}{\rightsquigarrow} (\mathcal{P}(y), T)$ for $x, y \in \Gamma$, such that $j(x) = \overset{\circ}{E}$ and $j(y) = \overset{\circ}{T}$. We need to show existence of the transition $x \overset{*}{\rightharpoonup} y$. The assumption $(\mathcal{P}(x), E) \overset{*}{\rightsquigarrow} (\mathcal{P}(y), T)$ means that there exists a sequence of causal relationships

$$cr_1, ..., cr_{n-1}, cr_i : \; \epsilon_i \;\; \textit{causes} \;\; \rho_i \;\; \textit{if} \;\; \Phi_i,$$

underlying the propagation, such that the relationship cr_i is applicable to the pair (q_i, E_i) yielding the pair (q_{i+1}, E_{i+1}), for $1 \leq i < n$, and $q_1 = \mathcal{P}(x)$, $q_n = \mathcal{P}(y)$, $E_1 = E$, $E_n = T$.

The required proof is also obtained by simple induction. For the case $n = 2$, the lemma's assumptions are $E_1 \subseteq \mathcal{P}(x)$ and $(\mathcal{P}(x), E_1) \rightsquigarrow (\mathcal{P}(y), E_2)$. By Definition 5.2, there exists an underlying causal relationship $cr_1 : \epsilon_1$ *causes* ρ_1 *if* Φ_1, such that

$\Phi_1 \wedge \neg\rho_1$ is true in $\mathcal{P}(x)$, and $\epsilon_1 \in E_1$. The condition $E_1 \subseteq \mathcal{P}(x)$ ensures that ϵ_1 is also in $\mathcal{P}(x)$, hence

$$\mathcal{P}(x) \vdash \epsilon_1 \wedge \Phi_1 \wedge \neg\rho_1$$

Given $\epsilon_1 \in E_1$ and $j(x) = \overset{\circ}{E}_1$, we obtain

$$j(x) \vdash \overset{\circ}{\epsilon}_1$$

Application of the causal relationship yields the pair $(\mathcal{P}(y), E_2)$, where $j(y) = \overset{\circ}{E}_2$. Hence,

$$\mathcal{P}(y) = (\mathcal{P}(x) \setminus \{\neg\rho_1\}) \cup \{\rho_1\}$$

$$j(y) = (j(x) \setminus \{\neg\rho_1\}) \cup \{\rho_1\}.$$

According to Definition 7.4 of the relation $\rightharpoonup$, the established four facts show that $x \rightharpoonup y$.

The case case of length k is similar. Let $(\mathcal{P}(x_1), E_1) \overset{*}{\rightsquigarrow} (\mathcal{P}(x_k), E_k)$ and $x_1 \overset{*}{\rightharpoonup} x_k$, and assume $(\mathcal{P}(x_k), E_k) \rightsquigarrow (\mathcal{P}(x_{k+1}), E_{k+1})$ for the states $x_k, x_{k+1} \in \Gamma$, such that $j(x_k) = \overset{\circ}{E}_k$ and $j(x_{k+1}) = \overset{\circ}{E}_{k+1}$. Lemma 5.5 establishes that $E_k \subseteq \mathcal{P}(x_k)$ and ensures that ϵ_k is also in $\mathcal{P}(x_k)$. This and the argument analogous to the $n = 2$ case show that $x_k \rightharpoonup x_{k+1}$, and transitivity of $\rightharpoonup$ yields that $x_1 \overset{*}{\rightharpoonup} x_{k+1}$. By induction, the transition $x \overset{*}{\rightharpoonup} y$ exists. □

Lemma 7.7 For a state $w \in \mathcal{W}$, there exists an ordering $\ll$ with respect to the mirror information state $\mu(w)$, such that for any action law $\langle C, a, E\rangle$,

$$\{\|E\|_w\} = min(\ll_{\mu(w)}, [E]^{\Gamma}).$$

Proof To prove this lemma we construct the required ordering $\ll$. We base the preferential structure in the information space Γ on the PMA ordering. First of all, given two information states γ_1 and γ_2 that belong to *different* information-neighbourhoods (i.e., have different projections), we would consider that γ_1 is closer to β than γ_2, denoted $\gamma_1 \ll_\beta \gamma_2$, if $\mathcal{P}(\gamma_1)$ is closer to $\mathcal{P}(\beta)$ than $\mathcal{P}(\gamma_2)$ in terms of the PMA ordering: $\mathcal{P}(\gamma_1) \prec_{\mathcal{P}(\beta)} \mathcal{P}(\gamma_2)$. Similarly, if the corresponding projections are not comparable: neither $\mathcal{P}(\gamma_1) \prec_{\mathcal{P}(\beta)} \mathcal{P}(\gamma_2)$ nor $\mathcal{P}(\gamma_2) \prec_{\mathcal{P}(\beta)} \mathcal{P}(\gamma_1)$, we do not specify any preference between γ_1 and γ_2. Secondly, we shall define a preference between the states γ_1 and γ_2 when they are both projected onto the same state in $\mathcal{W}$, more precisely, $\mathcal{P}(\gamma_1) = \mathcal{P}(\gamma_2)$. To do this we need a few auxiliary functions.

We define the *observed change* between two normal states w and q, denoted $q \overset{\circ}{-} w$, as the set of literals in $(q \setminus w)$ expressed in terms of justifier literals. More precisely,

$$q \overset{\circ}{-} w \;=\; \overset{\circ}{U}, \quad \text{where } \; U = q \setminus w.$$

Put simply, the observed change is the literals f from the state q that are not present in the state w, expressed in terms of justifier literals $\overset{\circ}{f}$. For example, given two states $q = \{a, \neg b, c\}$ and $w = \{a, b, c\}$, the observed change is $q \overset{\circ}{-} w = \{\neg \overset{\circ}{b}\}$.

The observed change can now be compared with the *justified* change – the change reflected by the justifier literals. To do this we use the symmetric difference $\mathit{Diff}(x, y) = (y \setminus x) \cup (x \setminus y)$. More precisely, we define the *divergent change*, $\triangle(\gamma, \beta)$, for two information states γ and β, as the set of justifier literals that appear either in the observed change set $\mathcal{P}(\gamma) \overset{\circ}{-} \mathcal{P}(\beta)$ or among the justifier literals $j(\gamma)$, but not in both. More precisely,

$$\triangle(\gamma, \beta) \;=\; \mathit{Diff}(\, \mathcal{P}(\gamma) \overset{\circ}{-} \mathcal{P}(\beta),\;\; j(\gamma)\,).$$

For example, if $\beta = \{a, b, c\}$ and $\gamma_1 = \{a, \neg b, c, \neg \overset{\circ}{b}\}$, then $\triangle(\gamma_1, \beta) = \emptyset$, and for $\gamma_2 = \{a, \neg b, c, \neg \overset{\circ}{a}, \neg \overset{\circ}{b}\}$ we obtain $\triangle(\gamma_2, \beta) = \{\neg \overset{\circ}{a}\}$.

Now we are ready to construct the preferential structure on the information space. The set $\mathcal{O}$ is a set of orderings $\ll$ defined on information states in such a way that respective projections satisfy the PMA ordering, while preferring subsets with smaller divergent change within each information-neighbourhood. More precisely,

$$\gamma_1 \ll_\beta \gamma_2 \text{ if and only if}$$

either $\mathcal{P}(\gamma_1) \prec_{\mathcal{P}(\beta)} \mathcal{P}(\gamma_2)$ or (both $\mathcal{P}(\gamma_1) = \mathcal{P}(\gamma_2)$ and $\triangle(\gamma_1, \beta) \subseteq \triangle(\gamma_2, \beta)$).

In other words, an information state γ_1 is nearer to β than γ_2 if and only if

– The states γ_1 and γ_2 correspond to different neighbourhoods, and the projection of γ_1 to normal space is nearer to the projection of β than the projection of γ_2, or

– If the projections are the same and the observed change in γ_1 diverges from the justified change at most as much as in γ_2.

This two-tiered PMA-based preference relation orders information states from different neighbourhoods, and then prefers subsets with smaller divergent change within each neighbourhood. Now we demonstrate that the constructed orderings have the desired property:

$$\{\|E\|_w\} = \mathit{min}(\ll_{\mu(w)}, [E]^\Gamma).$$

Consider $\prec_w$-minimal states among the post-condition states $[E]$. There is only one such minimal state because the post-condition E is expressed as a consistent set of literals. By definition, $\mathcal{P}(\|E\|_w)$ is the $\prec_w$-minimal state in $[E]$, and hence, there is no other state $p \in [E]$ such that $p \prec_w \mathcal{P}(\|E\|_w)$. Therefore, for any information state γ in $[E]^\Gamma$, such that $\mathcal{P}(\gamma) \neq \mathcal{P}(\|E\|_w)$, we obtain $\mathcal{P}(\|E\|_w) \prec_w \mathcal{P}(\gamma)$.

Since $w = \mathcal{P}(\mu(w))$, the preference $\mathcal{P}(\|E\|_w) \prec_{\mathcal{P}(\mu(w))} \mathcal{P}(\gamma)$ holds as well. It follows then, from the definition of the preference relation $\ll$ that the state $\|E\|_w$ is nearer to $\mu(w)$ in terms of $\ll_{\mu(w)}$ than any state γ from any other information-neighbourhood in $[E]^\Gamma$.

This leaves only those information space contenders for minimality that belong to the same information-neighbourhood as $\|E\|_w$. Among them, however, the state $\|E\|_w$ would be minimal because the divergent change between $\mu(w)$ and $\|E\|_w$ is empty: $\triangle(\|E\|_w, \mu(w)) = \emptyset$. To verify this, we note that $\mathcal{P}(\|E\|_w) \overset{\circ}{-} \mathcal{P}(\mu(w)) = \mathcal{P}(\|E\|_w) \overset{\circ}{-} w = \overset{\circ}{E}$ and $j(\|E\|_w) = \overset{\circ}{E}$. Other information states in the neighbourhood differ from $\|E\|_w$ at least in one justifier literal, making their divergent change sets non-empty. Therefore,

$$\{\|E\|_w\} = min(\ll_{\mu(w)}, [E]^{\Gamma}). \qquad \square$$

Theorem 7.8 *For every action system based on causal relationships there exists a selection-equivalent action system* $\langle \mathcal{W}, \mathcal{D}, \Gamma, \mathcal{E}, \mathcal{O}, \mathcal{M} \rangle$*.*

Conversely, for every action system $\langle \mathcal{W}, \mathcal{D}, \Gamma, \mathcal{E}, \mathcal{O}, \mathcal{M} \rangle$ *there exists a selection-equivalent action system based on causal relationships.*

Proof ($\Longrightarrow$) This part of the theorem follows from the construction,

- Specifying $\mathcal{M} = \rightharpoonup$;
- Including in the set $\mathcal{E}$ all the actions a from the set of action laws A, where each action law has the form $\langle C, a, E \rangle$, E being the direct effect of a;
- Choosing the set $\mathcal{D}$ as a subset of $\mathcal{W}$ such that its elements satisfy the constraints D;
- Defining the projection function $\mathcal{P}$ as in Definition 7.2, i.e., as the function mapping an information state $s \in \Gamma$ to a state $r \in \mathcal{W}$, such that $r = s \cap L_{\mathcal{F}}$;
- Constructing the orderings $\ll$ in the preferential structure $\mathcal{O}$ as specified in Lemma 7.7.

Given the presented construction and Lemma 7.7, the selection function

$$Res(w, a) = \\ \mathcal{D} \cap [E] \cap \mathcal{X}(\{\rho \in \mathcal{K}_{\mathcal{M}} : \ \mathcal{M}^*(\varphi, \rho), \ \text{where}\varphi \in min(\ll_{\mu(w)}, [E]^{\Gamma})\}).$$

can be re-written as

$$Res(w, a) = \mathcal{D} \cap [E] \cap \mathcal{X}(\{\rho \in \mathcal{K}_{\mathcal{M}} : \ \|E\|_w \overset{*}{\rightharpoonup} \rho\}).$$

As usual, E denotes the direct effect of the action law $\langle C, a, E \rangle$. In order to show the desired selection-equivalence, we need to show that a) $Res(w,a) \subseteq Res^*_{R,D,\mathcal{L}}(w, a)$ and b) $Res^*_{R,D,\mathcal{L}}(w, a) \subseteq Res(w, a)$.

a) Let $r \in Res(w, a)$. In other words, $r \in \mathcal{D} \cap [E]$ and $r \in \mathcal{X}(\{\rho \in \mathcal{K}_{\mathcal{M}} : \|E\|_w \overset{*}{\rightharpoonup} \rho\})$. The latter inclusion can be simplified as $r = \mathcal{P}(\rho)$, where $\rho \in \mathcal{K}_{\mathcal{M}}$ and $\|E\|_w \overset{*}{\rightharpoonup} \rho$. By definition of the trigger state, $\mathcal{P}(\|E\|_w) = (w \setminus C) \cup E$, and hence, $E \subseteq \mathcal{P}(\|E\|_w)$. Then, using Lemma 7.5, $\|E\|_w \overset{*}{\rightharpoonup} \rho$ implies $(\mathcal{P}(\|E\|_w), E) \overset{*}{\rightsquigarrow} (\mathcal{P}(\rho), T)$, where $\overset{\circ}{E} = j(\|E\|_w)$ by definition of the trigger state, and $\overset{\circ}{T} = j(\rho)$. This means that $((w \setminus C) \cup E, E) \overset{*}{\rightsquigarrow} (r, T)$ for some T. This condition, together with $r \in \mathcal{D} \cap [E]$, implies that $r \in Res^*_{R,D,\mathcal{L}}(w, a)$, and hence, $Res(w, a) \subseteq Res^*_{R,D,\mathcal{L}}(w, a)$.

b) Let $r \in Res^*_{R,D,\mathcal{L}}(w, a)$. In other words, $((w \setminus C) \cup E, E) \overset{*}{\rightsquigarrow} (r, T)$ for some T, and $r \in \mathcal{D} \cap [E]$. By definition of the trigger state, $\mathcal{P}(\|E\|_w) = (w \setminus C) \cup E$. Therefore, $(\mathcal{P}(\|E\|_w), E) \overset{*}{\rightsquigarrow} (\mathcal{P}(\rho), T)$, where $r = \mathcal{P}(\rho)$, and $\overset{\circ}{T} = j(\rho)$. Also, $E \subseteq \mathcal{P}(\|E\|_w)$. Lemma 7.5 immediately yields $\|E\|_w \overset{*}{\rightharpoonup} \rho$. Moreover, $\rho \in \mathcal{K}_{\mathcal{M}}$ because its projection, the state r, is the end point of causal propagation –

otherwise, there would be more causal relationships to apply, and r would not be in $Res^*_{R,D,\mathcal{L}}(w,a)$. We obtained that $r \in \mathcal{D} \cap [E]$ and $r = \mathcal{P}(\rho)$, where $\rho \in \mathcal{K}_{\mathcal{M}}$ and $\|E\|_w \overset{*}{\rightharpoonup} \rho$. This means that $r \in Res(w,a)$, and hence, $Res^*_{R,D,\mathcal{L}}(w,a) \subseteq Res(w,a)$.

The proofs a) and b) together establish that $Res^*_{R,D,\mathcal{L}}(w,a) = Res(w,a)$.

($\Longleftarrow$) The second part of the theorem is essentially trivial. Let $\mathcal{M}'$ denote a *stratified* causal relation, extracted from a given relation $\mathcal{M}$ by preserving only links connecting to stable states: $\mathcal{M}'(p,q)$ if and only if $\mathcal{M}^*(p,q)$ and $q \in \mathcal{K}_{\mathcal{M}}$. We parse the stratified binary relation $\mathcal{M}'$ into fully-qualified causal relationships. In other words, for every pair of information states such that $\mathcal{M}^*(\alpha,\beta)$ and $\beta \in \mathcal{K}_{\mathcal{M}}$ we consider the pair of states $x = \mathcal{P}(\alpha)$ and $y = \mathcal{P}(\beta)$. Then we create m^2 causal relationships

$$\epsilon_i \text{ } \textit{causes} \text{ } \rho_j \text{ } \textit{if} \text{ } \epsilon_1 \wedge \ldots \wedge \epsilon_i \wedge \ldots \wedge \epsilon_m,$$

where $x = \{\epsilon_1, \ldots, \epsilon_i, \ldots, \epsilon_m\}$ and $y = \{\rho_1, \ldots, \rho_j, \ldots, \rho_m\}$. The action system based on these causal relationships produces the same successor states as $Res(w,a)$. □

References

1. Bunge, M.: Causality. The Place of the Causal Principle in Modern Science. Harvard University Press, Cambridge, Massachusetts (1959)
2. Bunge, M.: Treatise on Basic Philosophy. Ontology I: The Furniture of the World. D. Reidel Publishing Company, Dordrecht-Holland/Boston-USA (1977)
3. Davidson, D.: Causal relations. In: Sosa, E., Tooley, M. (eds.) Causation. Oxford University Press, New York, USA (1993)
4. Ducasse, C.J.: On the nature and the observability of the causal relation. In: Sosa, E., Tooley, M. (eds.) Causation. Oxford University Press, New York, USA (1993)
5. Eco, U.: The Name of the Rose. Minerva, London (1992)
6. Fikes, R., Nilsson, N.J.: STRIPS: a new approach to the application of theorem proving to problem solving. Artif. Intell. **2**, 189–208 (1971)
7. Galton, A.: Time and change for AI. In: Gabbay, D.M., Hogger, C.J., Robinson, J.A. (eds.) Epistemic and Temporal Reasoning. Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 4, Clarendon, Oxford (1995)
8. Geffner, H.: Causality, constraints and the indirect effects of actions. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence, Aichi, Japan, 555–560 (1997)
9. Ginsberg, M.L., Smith, D.E.: Reasoning about action I: a possible worlds approach. Artif. Intell. **35**, 165–195 (1988)
10. Giunchiglia, E., Lifschitz, V.: Dependent fluents. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1964–1969 (1995)
11. Giunchiglia, E., Lee, J., Lifschitz, V., McCain, N., Turner, H.: Nonmonotonic causal theories. Artif. Intell. **153**(1–2), 49–104 (2004)
12. Kartha, G.N., Lifschitz, V.: Actions with indirect effects. In: Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning, Bonn, Germany, 341–350 (1994) (preliminary report)
13. Lifschitz, V.: Computing circumscription. In: Proceedings of the 9th International Joint Conference on Artificial Intelligence, Los Angeles, CA, USA, 121–127 (1985)
14. Lifschitz, V.: Two components of an action language. Ann. Math. Artif. Intell. **21**(2–4), 305–320 (1997)
15. Lin, F.: Embracing causality in specifying the indirect effects of actions. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1985–1991 (1995)
16. Mackie, J.L.: The Cement of the Universe. Oxford University Press, London, UK (1974)

17. Mackie, J.L.: Causes and conditions. In: Sosa, E., Tooley, M. (eds.) Causation. Oxford University Press, New York, USA (1993)
18. McCain, N., Turner, H.: A causal theory of ramifications and qualifications. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1978–1984 (1995)
19. McCain, N., Turner, H.: Causal theories of action and change. In: Proceedings of the 14th National Conference on Artificial Intelligence and 9th Innovative Applications of Artificial Intelligence Conference, Providence, RI, 460–465 (1997)
20. McCarthy, J., Hayes, P.: Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer, B., Michie, D. (eds.) Machine Intelligence IV, pp. 463–502. Elsevier, New York (1969)
21. McCarthy, J.: Circumscription – a form of non-monotonic reasoning. Artif. Intell. **13**, 27–39 (1980)
22. McIlraith, S.A.: Integrating actions and state constraints: a closed-form solution to the ramification problem (sometimes). Artif. Intell. **116**(1–2), 87–121 (2000)
23. Mellor, D.H.: The Facts of Causation. Routledge, London, New York, UK (1995)
24. Mellor, D.H.: Real Time II. Routledge, London, New York, UK (1998)
25. Pagnucco, M., Peppas, P.: Causality and minimal change demystified. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, USA, 125–130 (2001)
26. Peppas, P.: Belief change and reasoning about action. An axiomatic approach to modelling inert dynamic worlds and the connection to the logic of theory change. Ph.D. thesis, University of Sydney (1993)
27. Peppas, P., Pagnucco, M., Prokopenko, M., Nayak, A., Foo, N.: Preferential semantics for causal systems. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 118–123 (1999)
28. Prokopenko, M., Pagnucco, M., Peppas, P., Nayak, A.: Causal propagation semantics – a study. In: Proceedings of the 12th Australian Joint Conference on Artificial Intelligence, Sydney, Australia, 378–392 (1999)
29. Prokopenko, M., Pagnucco, M., Peppas, P., Nayak, A.: A unifying semantics for causal ramifications. In: Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence, Melbourne, Australia, 38–49 (2000)
30. Prokopenko, M., Pagnucco, M., Peppas, P., Nayak, A.: Capturing context in causal propagation. In: Proceedings of the IJCAI-01 Workshop on Nonmonotonic Reasoning, Action and Change, Seattle, USA, 95–102 (2001)
31. Prokopenko, M.: A unifying semantics for causal reasoning about action. Ph.D. thesis, Macquarie University (2002)
32. Reiter, R.: A logic for default theory. Artif. Intell. **13**, 81–132 (1980)
33. Sandewall, E.: Features and Fluents. Oxford University Press, New York, USA (1994)
34. Sandewall, E., Shoham, Y.: Non-monotonic temporal reasoning. In: Gabbay, D.M., Hogger, C.J., Robinson, J.A. (eds.) Epistemic and Temporal Reasoning. Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 4, Clarendon, Oxford (1995)
35. Sandewall, E.: Assessments of ramification methods that use static domain constraints. In: Proceedings of the 5th International Conference on Knowledge Representation and Reasoning, Cambridge, Massachusetts, 99–110 (1996)
36. Shanahan, M.: The ramification problem in the event calculus. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 140–146 (1999)
37. Shoham, Y.: Chronological ignorance: experiments in nonmonotonic temporal reasoning. Artif. Intell. **36**, 279–331 (1988)
38. Shoham, Y.: Reasoning About Change. MIT Press, Cambridge, Massachusetts (1988)
39. Thielscher, M.: Computing ramification by postprocessing, In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1994–2000 (1995)
40. Thielscher, M.: Ramification and causality. Artif. Intell. **89**, 317–364 (1997)
41. Thielscher, M.: How (not) to minimize events. In: Proceedings of the 6th International Conference on Knowledge Representation and Reasoning, Trento, Italy, 60–73 (1998)
42. Tooley, M.: Causation. A Realist Approach. Clarendon, Oxford (1987)
43. Turner, H.: Representing actions in logic programs and default theories: a situation calculus approach. J. Log. Program. **31**(1–3), 245–298 (1997)
44. von Wright, G.H.: On the logic and epistemology of the causal relation. In: Sosa, E., Tooley, M. (eds.) Causation. Oxford University Press, New York, USA (1993)
45. Winslett, M.: Reasoning about actions using a possible models approach. In: Proceedings of the 7th National Artificial Intelligence Conference, Saint Paul, MN, 89–93 (1988)