A Dynamical Systems Approach to Causation

Peter Fazekas[1,2,*], Balázs Gyenis[3,4],
Gábor Hofer-Szabó[4] and Gergely Kertész[5]

[1] Centre for Philosophical Psychology, University of Antwerp
[2] Center of Functionally Integrative Neuroscience, Aarhus University
[3] Department of Philosophy, Logic and Scientific Method, LSE
[4] Institute of Philosophy, Hungarian Academy of Sciences
[5] Department of Philosophy, Durham University

[*] corresponding author
fazekas.peter@gmail.com

Abstract

Our approach aims at accounting for causal claims in terms of how the physical states of the underlying dynamical system evolve with time. Causal claims assert connections between two sets of physicals states—their truth depends on whether the two sets in question are genuinely connected by time evolution such that physical states from one set evolve with time into the states of the other set. We demonstrate the virtues of our approach by showing how it is able to account for typical causes, causally relevant factors, being 'the' cause, and cases of overdetermination and causation by absences.

Keywords: causation; physical causation, folk causation, dynamical systems; state space; time evolution.

1

1. Introduction

In the 1974 classic action drama, *The Towering Inferno*, the overworked wiring of the Glass Tower (a 138-story skyscraper) develops a short-circuit in a storage room on the night of the opening ceremony turning the colossal building into a huge towering inferno. The storage room is filled with combustible material, while the oxygen level of the air is normal. So when the wiring develops a short-circuit and starts sparking, the combustible material catches fire and the flames sweep through the building. The occurrence of the short-circuit, in the presence of combustible material and oxygen, causes the burning down of the Glass Tower.

Our goal in this paper is to propose a novel way to analyse such causal claims.[1] Typically, philosophical approaches to causation either concentrate on deciphering everyday causal parlance, i.e. analysing the truth conditions of causal claims, or they try to capture what objective relations there might be (as described by our best scientific theories) that could underlie causal statements (cf. *conceptual* vs. *empirical* analysis—Dowe, 2000).[2] The novel framework we develop in this paper follows the second route. Our main objective is to account for causal claims and explain causal intuitions in terms of an approach to causal systems that is inspired by, and can hopefully be seamlessly integrated with, how physical theories characterise similar systems.

Emphasising the importance of system-level thinking is an inherent part of our proposal. As it happens, there are many different physical accounts of causation already on the table (Fair, 1979; Salmon, 1984; Dowe, 2000; Kistler, 2006) that all focus on features of individual interactions like transmission of certain marks (e.g. Salmon, 1984) or transference of conserved quantities (e.g. Fair, 1979). However, the scenarios described both by everyday and scientific causal claims (cf. e.g. 'sleep deprivation causes hallucinations') are often rather complex such that the possibility of decomposing them into sets of individual interactions is clearly out of sight. These scenarios are more like the systems statistical mechanics, chaos theory, control theory or cognitive science focus on: complex and dynamic.

---

[1] Depending on whether causes and effects are characterised by specific token level or by general type level descriptions our framework is equally able to account for both token and type causation (Sec. 2.2).

[2] Counterfactual (Lewis, 1986), and other difference making accounts (Woodward, 2003) follow the first route, whereas physical accounts of causation (Fair, 1979; Salmon, 1984; Dowe, 2000; Kistler, 2006) are more concerned about the second.

Our analysis of causal systems relies on the same toolkit that these scientific approaches to studying different forms of complex dynamical systems utilise—namely the characterisation of the system's behaviour in the space of all of its possible states.[3] Section 2 develops our account in detail and demonstrates how the truth of causal claims can be grounded by certain features of the underlying physical systems. Section 3 explains how key causal notions can be defined in a straightforward and intuitive way within our approach. Section 4 illustrates how classical cases that pose serious difficulties for existing accounts of causation (overdetermination and misconnection) can be uniformly and unproblematically accounted for in our framework. Finally, Section 5 concludes the paper.

2. The Proposal

In our initial example, the occurrence of the short-circuit in a room filled with combustible material and oxygen leads to a towering inferno. Behind this 'leads to' clause there is a nomic relation between physical states: a state of the world that co-instantiates the short-circuit, combustible material and oxygen evolves with time—according to the inherent dynamics of the system determined by laws of nature—into a state that instantiates the burning Glass Tower. So there is a fact about the physical world that contributes to rendering the claim that it was the short-circuit (in the presence of oxygen and combustible material) that caused the burning down of the skyscraper true: namely how certain physical states evolve with time.

However, this is not the only sort of fact that determines the truth of a causal claim. Different causal descriptions utilise different properties that carve up the space of possible physical states differently, and thus the set of physical states that co-instantiates them will be different. Therefore, the question is whether the actual set of properties invoked by a particular causal claim to characterise a cause and an effect is such that the

---

[3] Some have already tried to approach causation from the perspective of dynamical systems theory. The view that is closest to our proposal is Christopher Hitchcock's sketch of what he calls 'Laplacian causation' (Hitchcock, 2012, pp.46-51). Holly Andersen (2017) has recently published an account of causation that relied on dynamical systems and phase-space terminology. However, her aim was to develop an information-theoretic approach, thus the framework she offers is fundamentally different from ours. In a recent series of papers List and Pivato (2015, 2019) utilised a dynamical systems approach and applied it to a range of philosophical phenomena — but not to causation. Earlier List and Menzies (2009) analysed higher-level causation — but not in the framework of dynamical systems approach. Our paper unites the two programmatic aims: it combines higher-level causation with lower-level dynamical processes.

set of physical states that co-instantiates the properties characterising the alleged cause evolves into the set of physical states that co-instantiates the properties characterising the effect. The truth of causal claims, thus, is determined by two sorts of facts: (1) the time evolution of physical states, and (2) whether and how causal claims pick those physical states out that are connected by time evolution.

To put it in another way, we propose that causal claims try to capture relationships between different sets of physical states dictated by how physical states evolve with time. A causal claim is true (in the first approximation) if the properties postulated by the claim carve up the space of possible physical states in such a way that they pick out related sets. This is the basic idea behind our account in a nutshell. In the following sub-sections we shall clarify the details of our account.

2.1 The time evolution of physical states

Causal systems are *dynamical systems in the minimal sense* that their physical states evolve with time in accordance with the systems' inherent dynamics. Such systems can be characterised by the space of all of their possible physical states plus a rule describing the time evolution of the system (see Fig. 1).
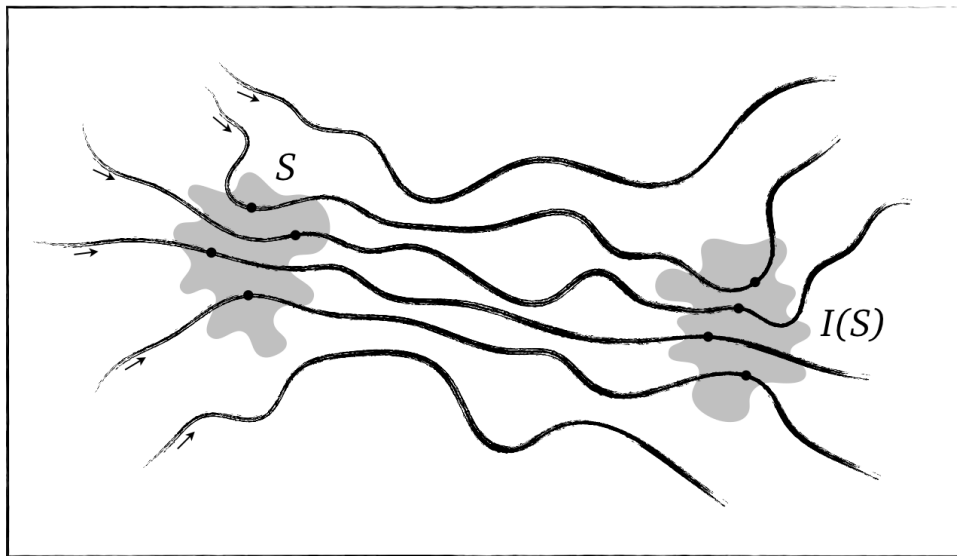

Figure 1.

A 2D projection of the state space. Arrows indicate the directions in which time evolution is driving physical states along corresponding trajectories. Physical states forming set *S* are taken by the system's dynamics (within a characteristic time bound) to the physical states forming set *I(S)*—that is, *S* is *pushed forward* into *I(S)* by time evolution. In other words, *I(S)* is the *image* of *S*.

The *space of all possible physical states* (for short: *state space)* is an abstract space of all the states that the system may find itself in, each point of which is a full characterisation of the entire system. The dimensions of this typically high-dimensional space are given by a minimal set of independent parameters, the *degrees of freedom*, that the system can fully be described with. Each point in this space—a *physical state*—corresponds to a specific value of all these parameters.

Physical states evolve with time: due to laws of nature the values of the parameters that are characteristic of a system change with time, and therefore a full characterisation of the system changes its position within the state space. The path along which a physical state moves in the abstract space as it evolves with time is called its *trajectory*. Laws of nature governing the properties of a system determine the *time evolution* of its physical states, and thus they determine the corresponding trajectories.[4]

How their physical states move with time in the state space is a fundamental feature of dynamical systems. Here we are interested in the phenomenon when a set of physical states is taken by time evolution into another set of physical states within a characteristic time bound (see Fn. 8). As we shall argue below, such a relationship between sets of physical states in the state space is a key component of why certain causal claims are true (Fig. 1).

Our approach is named after and greatly motivated by the theory of dynamical systems which is based on three key concepts, namely micro-state, time evolution and probability. The *par excellence* application of the theory is the statistical mechanical approach to classical thermodynamics. In its Boltzmannian version, one starts off with a set of macro-states supervening on the system's fine-grained micro-states. The challenge then is to explain macroscopic laws, such as the second law of thermodynamics, in terms of the microscopic behaviour of the system plus some probabilistic assumptions. In a sense, our account is a generalisation of this Boltzmannian model. We borrow the state space formalism, the time evolution and the micro-macro supervenience relation from this model. The 'macroscopic law' to be explained in our account is causation itself.

---

[4] Hence, the scope of our approach is limited to deterministic systems. This limitation does not necessarily exclude the application of our analysis to quantum mechanical systems, as there exist viable no-collapse interpretations (such as the Bohm–de Broglie theory) that render ordinary quantum mechanics deterministic in the relevant sense.

2.2 Causal descriptions and descriptive states

Everyday causal discourse, when asserting causal claims, relies on a set of *natural linguistic descriptions* (such as 'occurrence of short-circuit', 'presence of oxygen', and so on) to characterise causes, effects, and background conditions. Properties picked out by such descriptions correspond to sets of physical states instantiating the property in question. In other words, each property can be represented by a *region* of the state space.

Regions in the state space corresponding to independent properties cross-cut in all possible ways. So in the case of our example, there is a region consisting in physical states that co-instantiate the occurrence of a short-circuit, the presence of oxygen, and the presence of combustible material; there are regions that co-instantiate only two of the properties and are also defined by the negation, i.e. the non-instantiation of the third property (occurrence of a short-circuit and the presence of oxygen, but not the presence of combustible material); there are regions that instantiate only one of these properties; and there is a region that does not instantiate any of these properties (see Fig. 2).

Properties via the intersections of corresponding regions partition the state space. Individual elements (i.e. the sub-regions) of such partitioning correspond to complex states of affairs as they are characterised by the descriptions that a causal discourse invokes, hence we shall call them *descriptive states*. Descriptive states, thus, are determined by a list of which of the properties utilised in a particular causal discourse are instantiated and which are not.[5] From this perspective, then, causal claims assert certain connections between two descriptive states, the *cause state* and the *effect state*—two sets of physical states that are defined by the properties utilised by the causal claim to characterise the cause and the effect, respectively.

That is, causal claims provide macro-descriptions that correspond to regions of the state space defined by the set of independent parameters of the full micro-characterisations of the system a causal claim tries to tell us something about. The properties utilised by causal claims are macro-properties, such that many different physical states instantiate them. The set of physical states co-instantiating the properties utilised by a

---

[5] Of course, the number of relevant negative truths (conditions that do *not* hold in a particular situation) is always infinite. A descriptive state, however, is determined by only those properties that a particular causal discourse actually invokes. Introducing new properties into the characterisation of causes and effects results in a more fine-grained partitioning. Type-level descriptions provide coarser-grained partitioning, token-level descriptions provide finer-grained partitioning (since token-level descriptions are still macro descriptions—the partitioning they impose upon the state space still consists of regions).

causal claim to characterise the cause is called the cause state, whereas the set of physical states co-instantiating the properties characterising the effect is called the effect state. Cause and effect states are, thus, regions of the state space, and the causal claim asserts a connection between these two regions (represented by the dashed arrow in Fig. 2.).[6]
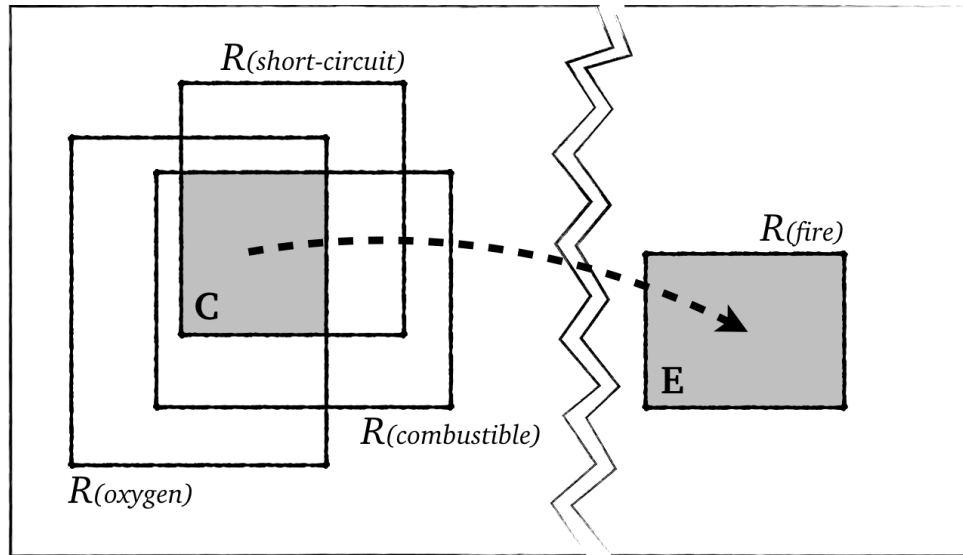

Figure 2.

A partitioning of the state space by the properties invoked by a causal claim. Properties correspond to regions (*R*), their intersections define descriptive states. Causal claims assert connections between two descriptive states, the cause state (**C**) and the effect state (**E**). This connection is indicated by the dashed line with arrowhead. *R(oxygen)*: 'presence of oxygen in the storage room of the Glass Tower at time t_1'; *R(combustible)*: 'presence of combustible material in the storage room of the Glass Tower at time t_1'; *R(short-circuit)*: 'occurrence of a short-circuit in the storage room of the Glass Tower at time t_1'; *R(fire)*: 'burning down of the Glass Tower at time t_2'; **C**: *R(combustible) ∩ R(oxygen) ∩ R(short-circuit)*—the cause state; **E**: *R(fire)*—the effect state.[7] Note that the effect state is indicated only in a simplified form, not correctly positioned relative to the cause state. The figure also disregards concrete trajectories. The double zigzag line emphasises these distortions. Bold capitals indicate descriptive states, italicised capitals indicate regions of the state space (i.e. sets of physical states).

2.3 Causal claims and projective states

What can establish the truth of certain causal claims asserting connections between different sets of physical states? As we have seen, there is a natural way in which physical

---

[6] Though developed independently, this aspect of our account resembles what Hitchcock calls 'region cause' (Hitchcock, 2012, p.49).

[7] The significance of indexing with time shall be discussed in Sec. 2.3, especially Fn. 8. See also Sec. 2.4 and Sec. 4.1.

states are connected in the state space: via time evolution. Time evolution takes a system from one physical state to another, and thus defines a trajectory within the phase space (see Sec. 2.1).

We propose that it is exactly this natural connection via time evolution that establishes the truth of certain causal claims. That is, what makes a causal claim true is how the physical states it picks out are related by time evolution: a causal claim is true if the (majority—see Sec. 2.4—of the) physical states of the cause region evolve with time into the physical states of the effect region. For example, the occurrence of a short-circuit, the presence of oxygen, and the presence of combustible material in one of its storage rooms causes, or 'leads to', the burning down of the Glass Tower if and only if the (majority of the) physical states co-instantiating the occurrence of a short-circuit, the presence of oxygen, and the presence of combustible material in one of the storage rooms evolve with time—within a characteristic time bound[8]—into the physical states instantiating the burning down of the Glass Tower.

Note that this condition is not fulfilled automatically: whether sets of physical states picked out by causal claims are connected by time evolution also depends on what kind of natural linguistic descriptions are utilised in any given causal discourse. Even if the state space of any given causal system is such that there are sets of physical states in it that evolve into other sets with time, causal discourse does not automatically postulate properties that correspond to such sets. Take, for example, the set of physical states co-instantiating the properties postulated by a causal discourse to characterise a particular cause. Time evolution pushes this set forward into another set of physical states, but this image set does not necessarily coincide with the set of physical states that co-instantiate the properties invoked by the causal discourse to characterise the effect (see Fig. 3). So the crucial question, which the fate of any causal claim hangs on, is whether the partitioning imposed upon the state space by the properties invoked by a particular causal discourse to describe alleged causes is able to pick out physical states

[8] Our everyday causal parlance is very sensitive to the temporal aspects of causal relations: it is not enough for a given state to evolve into another state within some finite time, it has to evolve into that within a reasonable time bound that is characteristic of the typical length of the causal influences that a particular causal discourse tries to describe. Individual physical states of a region might need slightly different amounts of time to evolve into another region, but these fall within an interval that corresponds to the characteristic time bound. For an illustration see the caption to Fig. 11.

that evolve within a characteristic time bound into physical states that instantiate the properties characterising the effect.
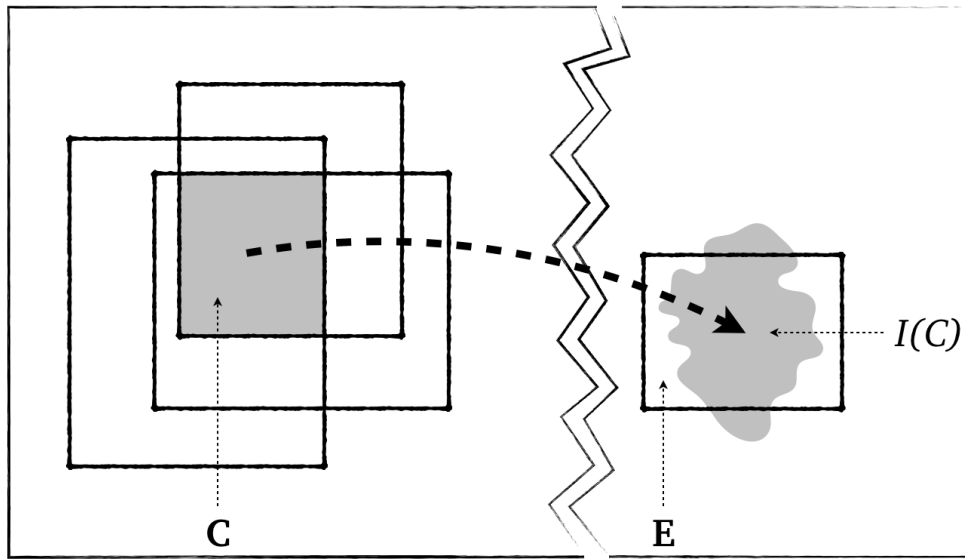

Figure 3.

Causal claims and time evolution. The asserted connection between descriptive states **C** and **E** (see Fig. 2) is grounded in the relation between the image *I(C)* of the set of physical states corresponding to **C** and the set of physical states corresponding to **E**. If the majority of *I(C)* is 'captured' by **E**, then the causal claim is true. (Note that here not all physical states of **C** evolve into **E**—see Sec. 2.4.)

From this perspective, then, the goal of the practitioners of causal discourse is to find descriptions that define descriptive states, i.e. an effect state and a corresponding cause state, such that the effect state can 'cover' or encompass the majority of the image of the cause state. In other words, causal claims need to be able to define causes and effects in such a way that ensures that there is 'systematic channelling' of trajectories from cause states to effect states, i.e. that the majority of the physical states of the cause region evolve into the effect region (see Fig. 3). In what follows, we shall call those descriptive states from which time evolution 'projects' into an effect state in this sense of channelling trajectories *projective states* of the effect state in question. The truth of a causal claim, thus, depends on whether the descriptive state corresponding to the alleged cause is indeed a projective state of the descriptive state corresponding to the particular effect.

2.4 Accuracy

Before moving on, a few more words need to be said about what exactly we mean by '*systematic* channelling', and why we need the '*majority*' clause in the definition of projective states. As a start, note that the tightness of fit, i.e. the *accuracy*, of the effect state around the image of the cause state—how accurately the effect state maps onto those physical states that are pushed forward by time evolution from the cause state—is of utmost importance.

To clarify the notion of accuracy, recall that whereas the question of how the image of a cause state develops—how time evolution pushes forward a region in the state space—is determined by physical facts (i.e. laws of nature and the inherent dynamics of the system), the related question of which physical states an effect state encompasses, is determined by facts about the actual causal discourse (i.e. the properties utilised to characterise states of affairs). Hence effect states and images of cause states are independent from each other, making it possible to adjust the size and position of an effect state around the image of a cause state by utilising more or less fine-grained (detailed) descriptions, thereby making the characterisation of the effect (of a given cause) more or less accurate.

On the one hand, note that utilising a too coarse-grained description the effect region might become so extended that it encompasses distinct images of different descriptive states (see Fig. 4). This can result in cases where two different descriptive states turn out to be projective states of the same effect state, because the effect state in question is defined in such broad terms that it maps onto two different sets of physical properties that are independently related (by time evolution) to the physical states of distinct descriptive states. That is, utilising too vague properties to characterise an effect might obscure the underlying structure that otherwise could have been uncovered by finer-grained descriptions. Importantly, when the causal discourse provides a finer-grained description of the possible causes of an effect together with a coarser-grained description of the effect itself, the use of the too vague properties to characterise the effect might create confusion with regard to the cause of the effect in question. In such a case, a finer-grained characterisation of the effect (matching the level of description

used to refer to possible causes) could allow for clarifying the situation by picking out the independent relations as distinct causal claims (see Fig. 4).[9]
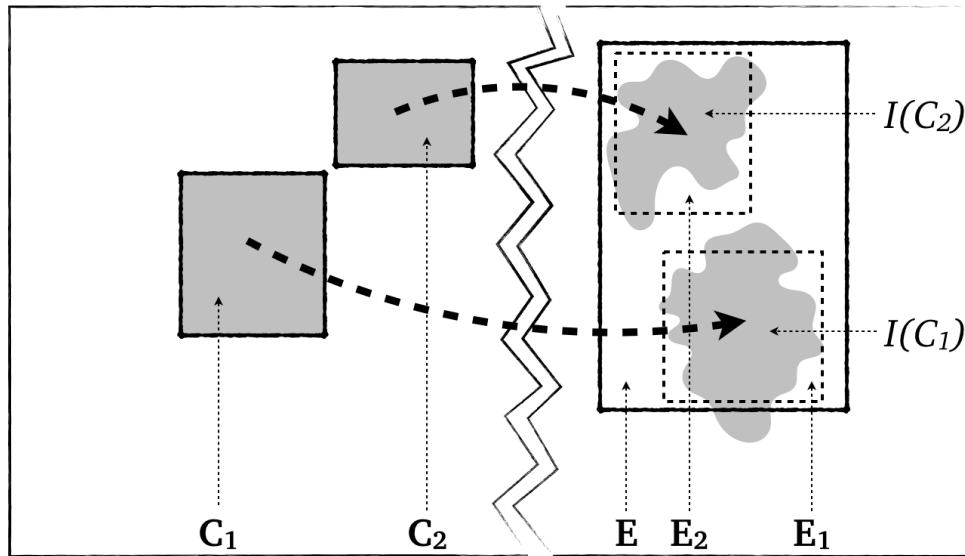

Figure 4.

Effect state **E** is defined in broad terms, and hence it maps onto a relatively large region of the state space encompassing two distinct sets of physical states, $I(C_1)$ and $I(C_2)$, that are images of two different descriptive states (**C₁** and **C₂**, respectively). This renders both descriptive state **C₁** and descriptive state **C₂** projective states of effect state **E**. However, with the aid of a more fine-grained description of the effect, defining distinct effect states **E₁** and **E₂**, the two independent relations, C_1-$I(C_1)$ and C_2-$I(C_2)$, could be picked out by two distinct causal claims (**C₁-E₁** and **C₂-E₂**).[10]

On the other hand, utilising a too fine-grained description of the effect resulting in a small effect region threatens with excluding too many physical states of the image of the cause state from the effect region, which can render alleged cause states non-projective states, since in this way the majority of the physical states of the cause state might not evolve into the small effect region (see Fig. 5).

Note that cases where all physical states co-instantiating the properties defining the cause state evolve into the effect state correspond to *exception-less* causal claims. If, however, there are physical states in the cause region not evolving into the effect region, then the causal claim in question will not be exception-less. Nevertheless, such a claim

---

[9] For a discussion of this issue in less abstract terms see Sec. 4.1.

[10] Cf. multiple realisability based considerations. If there is a descriptive state **C** encompassing **C₁** and **C₂** such that it is a projective state of **E**, then the causal claim **C-E** formulated in broader terms might be said to be multiply realised by causal claims **C₁-E₁** and **C₂-E₂** formulated in more fine-grained terms.

can still express a robust relationship depending on the proportion of the physical states that are taken into the effect state by time evolution.
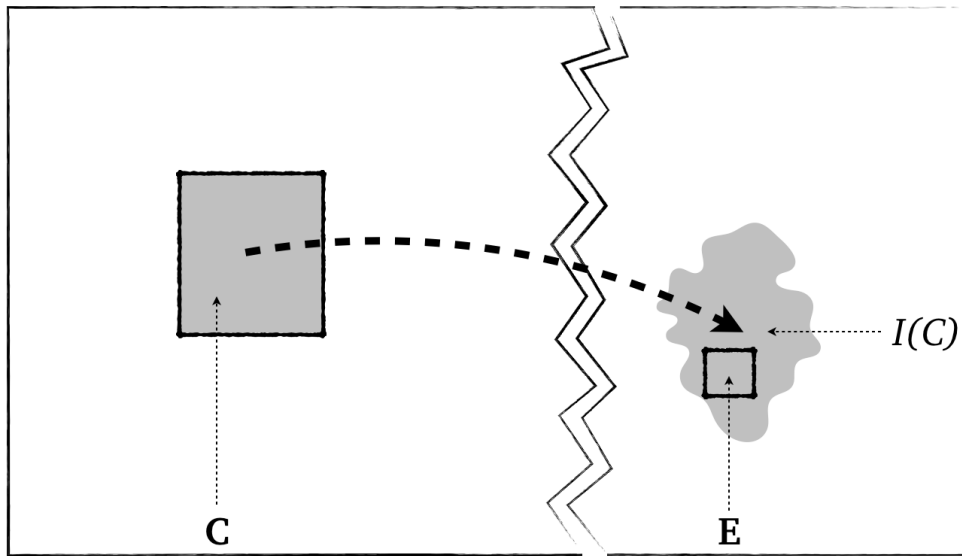

Figure 5.

A small effect state (**E**) defined by a fine-grained description of the effect. Many of the trajectories evolving from the cause state (**C**) never enter the effect state, which renders the cause state (**C**) a non-projective state of the particular effect state. Note that only the cause state itself is depicted here without the individual regions corresponding to the properties defining the cause state (compare Fig. 2).

Causal claims are rarely exception-less, therefore here we shall not require all of the physical states of the cause region to evolve into the effect region. Instead, we shall concentrate on how *robustness* could be warranted: the more physical states (the bigger proportion) of the cause region are taken into the effect region within a characteristic time bound, the more robust the causal claim can be.[11] It is this notion of robustness, thus, that the term '*systematic* channelling' tries to capture in the definition of projective states (see Sec. 2.3). Similarly, the 'majority' clause utilised to characterise projective states emphasises that typical causal claims are robust in the above sense. However, just as there is no objective criterion for how robust a causal claim should be, we leave the question of what proportion of the physical states of the cause region should evolve into

_____

[11] This notion of robustness captures the traditional understanding that robust causal relationships hold in a variety of backgrounds, as regions corresponding to different backgrounds intersect with the cause region cutting sub-regions out of the original cause region. If a causal relation doesn't hold in a given background, then the majority of the physical states from the corresponding sub-region will not evolve into the effect region.

the effect region open. Less robust causal claims are underlain by less systematic channelling.[12]

3. Causal Notions in Our Framework

Now that the fundamental components of our account are all on the table, let's see how they work together, how they allow us to interpret and explain causal notions that play central roles in causal parlance.

3.1 Typical causes

Though projective states are essential with regard to grounding causal claims, note that they are not unique: for any effect state there can be many different projective states. We have already seen why it might be so in the case of coarse-grained effect states (Sec. 2.4, Fig. 4). Consider, for example, a broad description of our original example ('burning down of a house'), and two different descriptive states (say, 'occurrence of a short-circuit in the presence of oxygen and combustible material', and 'throwing a burning match onto combustible material in the presence of oxygen') that both are projective states of the effect state in question.

In such cases, one can take into consideration all those trajectories that lead to the physical states of the effect, and check for the descriptive states crossed by them whether they are projective states of the effect region. Depending on what proportion of the trajectories leading to the effect cross any given projective state, the corresponding states-of-affairs will be more or less probable causes of the effect (in the sense of how probable it is that the actual effect has been caused by a particular state-of-affairs). That is, whereas the proportion of physical states of a projective state that evolve into the effect state determines the robustness of a causal claim, the proportion of the trajectories leading to the effect state that cross a given projective state determines the likelihood of the cause. The projective state that is crossed by the majority of the trajectories leading to the physical states of the effect (if there is such a projective state) might be called the *typical projective state* of the effect, and the co-instantiation of the corresponding properties might be called the *typical cause* of the effect.

---

[12] All these considerations, then, mutually constrain the sizes of the cause and effect regions, i.e. the level of detail in which causes and effects can be specified. Due to limitations in space, here we can only propose that this is what ultimately grounds intuitions about the proportionality of causation (Yablo, 1992).

3.2 Relevant causes and principal projective state

Besides alternative projective states on different trajectories,[13] there can be alternative projective states even along the same trajectory. For example, consider the following description that one might try to provide as a characterisation of why the Glass Tower has burned down: [in the presence of oxygen and combustible material] fire broke out from the storage room. The descriptive state defined by the properties invoked by this characterisation is just as much a projective state of the effect state as the descriptive state corresponding to the original 'occurrence of short-circuit inside the storage room in the presence of oxygen and combustible material' description (see Fig. 6). But are both of these projective states equally relevant in bringing about the effect?
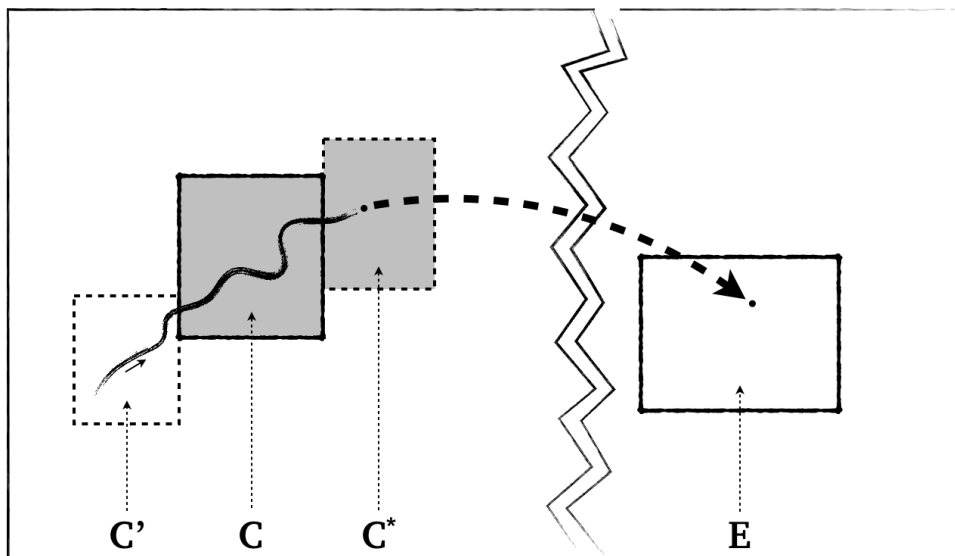
Figure 6.

Multiple projective states of the same effect state. Descriptive state **C** corresponds to the description 'occurrence of short-circuit in the storage room of the Glass Tower, in the presence of oxygen and combustible material, at time t_1'. Effect state **E** corresponds to the description 'the Glass Tower burns down at t_2'. Descriptive state **C*** corresponds to the description 'fire brakes out from the storage room of the Glass Tower, in the presence of oxygen and combustible material, at $t_1+\Delta t$' ($t_1 < t_1+\Delta t < t_2$). **C'** is a descriptive state corresponding to certain states-of-affairs at some $t_1-\Delta t$ (see main text). Both **C** and **C*** are projective states of **E**, whereas **C'** is not. Hence **C** is the principal projective state of **E**. Note that here only the actual trajectory is depicted.

Intuitively, they are not. As we get closer in time to the effect (the smoking ruin of the Glass Tower) one could provide more and more descriptions about how the fire

---

[13] We say that a descriptive state is *on* the backwards trajectory of a physical state iff it is crossed by the trajectory that evolves into said physical state.

spread through the building destroying bigger and bigger parts of it. However, what one is interested in (what, for instance, an investigator would be interested in) when looking for the cause of the burning down of the Glass Tower is finding what started the fire in the first place, i.e. finding a distinctive incident that somehow steered the course of further events towards the outcome (Strevens, 2004).

Since in our framework 'courses of events' correspond to trajectories, this idea of finding what 'steered the course of events towards the outcome' translates onto finding a region of the state space that plays a crucial role in the fate of the trajectories crossing it. In our terminology, this means finding a projective state of the effect region that is preceded by a descriptive state which is not a projective state. Following a trajectory leading to the effect region backwards from the physical state of the effect leads us through a series of projective states (that, for instance, in the case of our example, correspond to characterisations of what happens with the Glass Tower further and further away in the past of the instantiation of its smoking ruin). However, at one point the trajectory (still followed backwards) may enter into a descriptive state that is not projective (because the pre-short-circuit state of the Glass Tower is characterised in such a way that it does not pick out a set of physical states the majority of which evolve into the effect region). This transition between a non-projective descriptive state and a projective state corresponds to the distinctive incident that steered the course of events towards the outcome (to the extent that can be captured in the terms of the causal discourse). In our terminology, such projective state is the descriptive state where the projection starts, i.e. it is a difference making descriptive state that can capture the systematic channelling of

the trajectories towards the effect region. We call such a projective state *principal projective state*.[14]

That is, whereas the truth of a causal claim depends on whether the descriptive state defined by the properties invoked to characterise the alleged cause of an effect is a projective state of the corresponding effect region, the *relevance of the causal claim* depends on whether this projective state is the principal projective state. A *relevant cause* of an effect, thus, is the co-instantiation of the properties defining the principal projecting state.[15]

3.3 Causally relevant factors

As we have seen, not all projective states are equally relevant in characterising what contributes to bringing about a particular effect. In a certain sense, of course, all projective states of an effect state tell us something important about bringing about the effect: that the instantiation of the set of properties defining a projective state of the effect state is likely sufficient for bringing about the effect. However, it is a further question whether we can separate the *causally relevant factors* of bringing about the effect (i.e. the set of properties the instantiation of which has actually contributed to bringing about the actual instance of the effect) from *spurious causes* (i.e. extra factors that, though are co-instantiated with the causally relevant factors, in fact, do not contribute to bringing about the actual effect).

---

[14] Two further issues need to be emphasised here. First, the partitioning offered by a causal discourse is typically not specific enough for the entire time evolution of the trajectory in question, and thus it won't be able to carve regions out from the state space along the entire trajectory such that from these the majority of the points evolve into the effect state. Therefore, descriptive states that a trajectory—followed backwards—crosses are *not* always projective states, and thus for any given effect state there will probably be a corresponding principle projective state. Second, intuitively one may characterise the principal projective state as the first projective state into which the forward trajectory (that ends up in the effect state) enters after leaving a descriptive state that is not yet a projective state. Taking into account our restriction on characteristic time development this is another useful way to characterise our notion of principal projective state; however, strictly speaking, there is no guarantee that the principal projective state is unique. If one followed the trajectory further backwards it may be the case that after passing through several non-projective states the trajectory again enters into another (principal) projective state. Indeed when the dynamics is ergodic one would expect that the backwards trajectory would pass through again and again the same principal projective state. See also our discussion of the typical principal projective state in Sec. 4.2.

[15] In the literature, the notion of relevance is also used to distinguish between different descriptions of the same phenomenon (relying on different properties that are in a supervenience relation). In such cases, the different projective states cover the same section of a trajectory in different ways, but the trajectory enters and leaves these states at the same time (as supervenience is a synchronic relation). A full discussion of how our approach could be used in those cases requires a standalone paper. (See also the Conclusion.)

To see how causally irrelevant properties (that do not make a difference with regard to the occurrence of the effect in question) might play a part in defining a projective state, consider a twist on our original example. A few minutes before the short-circuit starts sparking, not aware of what is about to happen in one of the storage rooms, the organisers start the celebration dinner in the Promenade Room of the Glass Tower. In this case, the descriptive state defined by 'short-circuit in the presence of oxygen and combustible material, and guests dining'—which is partly defined by the property 'guests dining'—is a projective state of the burning down of the Glass Tower (see Fig. 7). Intuitively, however, the fact that the guests are dining is not causally relevant in bringing about the towering inferno.
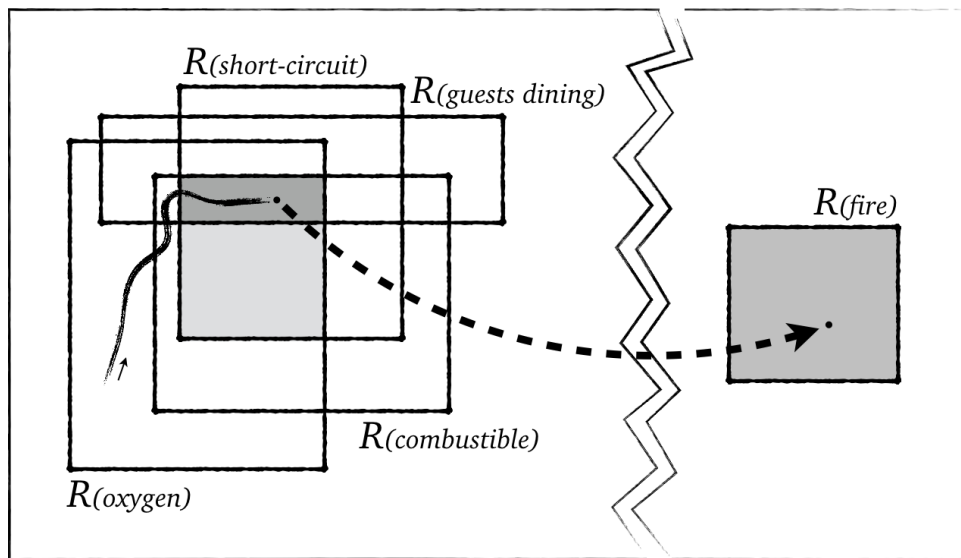

Figure 7.

Guests start dining right before the occurrence of the short-circuit in the storage room. The dark grey region on the cause side (corresponding to 'short-circuit in the presence of oxygen and combustible material, and guests dining') is a projective state of the effect. The light grey region (corresponding to 'short-circuit in the presence of oxygen and combustible material, and guests *not* dining') is also a projective state of the effect. The boundary of the region corresponding to the instantiation of 'guests dining' does not separate a projective state from a non-projective state.

Causally irrelevant properties that nevertheless partly define a projective state can easily be detected in our framework. As we have seen, regions in the state space corresponding to properties that are independent from each other intersect in all possible ways. So there will be descriptive states corresponding to all possible combinations of the instantiation and non-instantiation of the properties defining a given projective state.

Some of these descriptive states will also be projective states, others will not be (see Fig. 7). Those properties will be causally irrelevant whose instantiation (or non-instantiation, see 4.2) does not make a difference with regard to whether the descriptive state they partly define is a projective state or not. That is, in the case of causally irrelevant properties there will be no configuration of the instantiation or non-instantiation of the other properties defining the projective state in question for which the instantiation (or non-instantiation) of the causally irrelevant property could change whether the corresponding physical states evolve into the effect region or not.

Put it in another way, the boundaries of the regions corresponding to the instantiation of *causally relevant factors* (i.e. the properties that together form the minimal set of jointly likely sufficient properties of bringing about a particular effect) will at least sometime separate a projective state from a non-projective state. For causally irrelevant properties, this will never be true.

3.4 Background conditions and 'the' cause

In Sec. 3.2, we have seen that the most relevant causal claim is the one that relies on the principal projective state of an effect state. This is the state the occurrence of which made a difference with respect to bringing a particular effect about, and thus the co-instantiation of its non-accidental (i.e. causally relevant) defining properties might rightly be claimed to cause the occurrence of the effect.

However, casual talk is typically monocausal. Causal claims referring to *'the' cause* of an effect pervade both everyday and scientific causal discourse. In fact, causal intuitions forcefully select one of the contributing factors as 'the' cause of any given effect. This seems to be a fundamental feature of causal discourse: it most often classifies causally relevant factors into two categories: *background conditions* and *the cause*.

For example, in our test case our analysis so far suggests that the co-instantiation of combustible material, oxygen and a short-circuit led to the burning down of the Glass Tower. These properties form the set of causally relevant factors. But everyday intuitions do not stop here. Typical causal parlance refers to the short-circuit as 'the' cause of the fire while it renders the presence of oxygen and combustible material 'background conditions' (Hart and Honore, 1985). Is it possible to account for this distinction within our framework?

As it happens, our approach is readily able to pinpoint one of the properties relevantly defining the principal projective state of a particular instance of an effect as 'the' cause of the effect in question. Recall that the principal projective state is a projective state on a backwards trajectory from the effect state which is preceded by a descriptive state on the backwards trajectory that is not a projective state anymore. This is what allows us to say that the properties (relevantly, see Sec. 3.3) defining the principal projective state are the properties the co-instantiation of which makes a difference in bringing about the particular effect.

This train of thought naturally leads to a further conclusion. Trajectories, in order to enter the principal projective state, need to cross the boundary of one of the regions that define such a state. Regions defining descriptive states correspond to properties such that whereas the physical states outside the region do not, the physical states inside the region do instantiate the property in question. So inasmuch as entering the principal projective state 'determines the fate' of the trajectory, it is the instantiation of the very property corresponding to the region the boundary of which the trajectory crosses when it enters the principal projective state that ultimately does so.
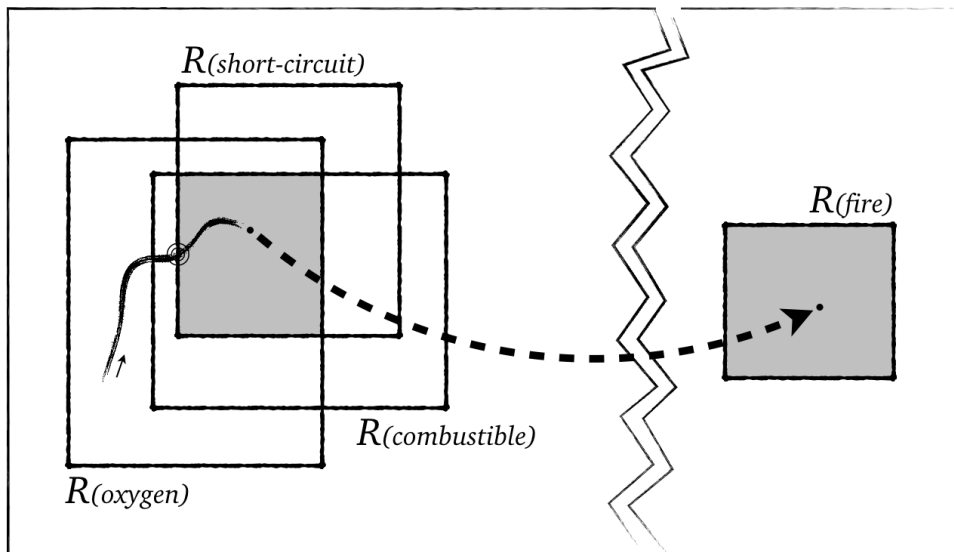

Figure 8.

"The" cause of the burning down of the Glass Tower. The double circle indicates that the trajectory enters the principal projective state of the effect by crossing the boundary of the region corresponding to the instantiation of the occurrence of the short-circuit.

Something important happens when a trajectory leading to a physical state of the effect enters the principal projective state. The descriptive state from which the trajectory enters the principal projective state is such that there is no systematic channelling of

trajectories from that descriptive state into the effect region. However, from the principal projective state there is systematic channelling. What makes this difference—what can possibly be responsible for this difference from the perspective of the causal description—is the only change that happens as the trajectory crosses the boundary between the non-projective state and the principal projective state, namely that one of the properties defining these descriptive states changes: it becomes or ceases to be instantiated. This change in the instantiation of this property, then, is *'the' cause* of the effect in question, whereas all the other (relevant) defining properties of the principal projective state will be background conditions (see Fig. 8).[16]

4. Applications

To demonstrate the strength of our approach we conclude the paper by applying our framework to two cases that pose serious problems for existing accounts of causation.

4.1 Overdetermination

Consider first overdetermination, i.e. a family of scenarios in which two or more causes are redundantly present, and compete, in a certain sense, for the very same effect. Overdetermination is famously problematic for counterfactual theories of causation since there is no counterfactual dependence between the competing redundant causes and the effect (Lewis, 1986). Similarly, it is also problematic for a Mackie-style analysis, since neither of the competing redundant causes are necessary for bringing about the effect, so neither of them is an INUS condition (Mackie, 1974).

In what follows we focus on a specific variety of overdetermination, namely *late preemption*, that—at first sight—looks problematic for our approach. (Other varieties of overdetermination, like symmetrical overdetermination and early preemption can straightforwardly be accounted for following the logic of our approach.) Fig. 9 is a representation of a modified version of Ned Hall's famous example. Billy and Suzy are throwing stones at a window. First, Billy throws his stone. Then Suzy also throws her

---

[16] In addition to such *actuality* consideration, causal selection can also be driven by *normative* (see Fn. 17) and *typicality* considerations (Hart and Honore, 1985). Due to limitations in space we can only indicate that besides typicality (see Sec. 3.1), the notions of *atypicality* and *abnormality*, which play a central role in selecting 'the' cause of an effect in typicality-driven cases, can also be accounted for within our approach. A full treatment of this issue shall follow in a subsequent paper.

stone, but she does so with a much greater thrust than Billy did, so her stone gets faster than Billy's, arrives to the window sooner than the boy's, and breaks it (Hall, 2004).
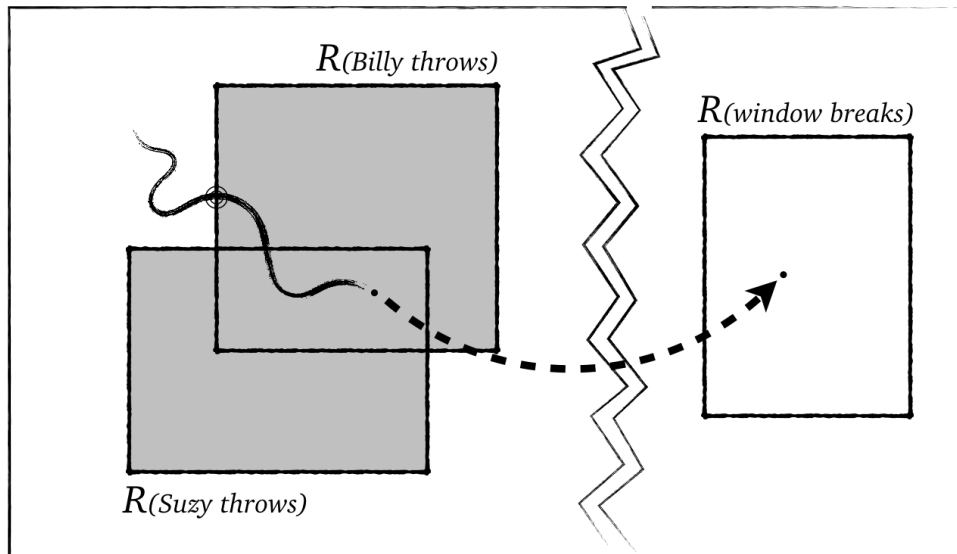

Figure 9.

Billy, Suzy and the window. Billy throws his stone, then Suzy throws hers. Suzy's stone gets ahead of Billy's and breaks the window. Greyed out regions indicate projective states of the breaking of the window. Our original methodology suggests that the principal projective state of the breaking of the window is Billy's throwing.

Even if counterfactual and INUS accounts are in trouble, everyday intuitions are not puzzled at all: since it was Suzy's stone that hit the window first and broke it, Suzy's throwing is the cause of the window's breaking. Can our framework account for everyday intuition in this case? Well, it seems that there is a problem. There are three projective states of the effect (the window breaks if only Billy throws, if only Suzy throws, and if they both throw), and our analysis picks out the one defined by Billy's throwing and Suzy's not throwing as the principal projective state (since that is the projective state the trajectory enters to from a non-projective state), and consequently Billy's throwing as 'the' cause of the effect.

To resolve this tension, recall what has been said about accuracy and converging trajectories in Sec. 2.4 and Sec. 3.1 (especially, see Fig. 4). A very general description of the effect can cover a broad region in the state space parts of which might be related by time evolution to very different physical states.

We propose that this is exactly what happens in the cases of late preemption. Consider Fig. 10, which is a modified version of Fig. 9 revealing that the original natural linguistic description picking out the effect (the window breaks) covers two distinct

subregions that are related by time evolution to different sets of physical states. It is possible to pick out these subregions of the region covered by the general description by more detailed descriptions, say 'the window breaks at t_1' and the 'the window breaks at t_2' ($t_1 < t_2$), such that different descriptive states will be projective states of these two new fine-grained effect states: the state when Billy throws but Suzy doesn't projects to the window breaking at t_2, whereas the states when Suzy throws project to the window breaking at t_1.
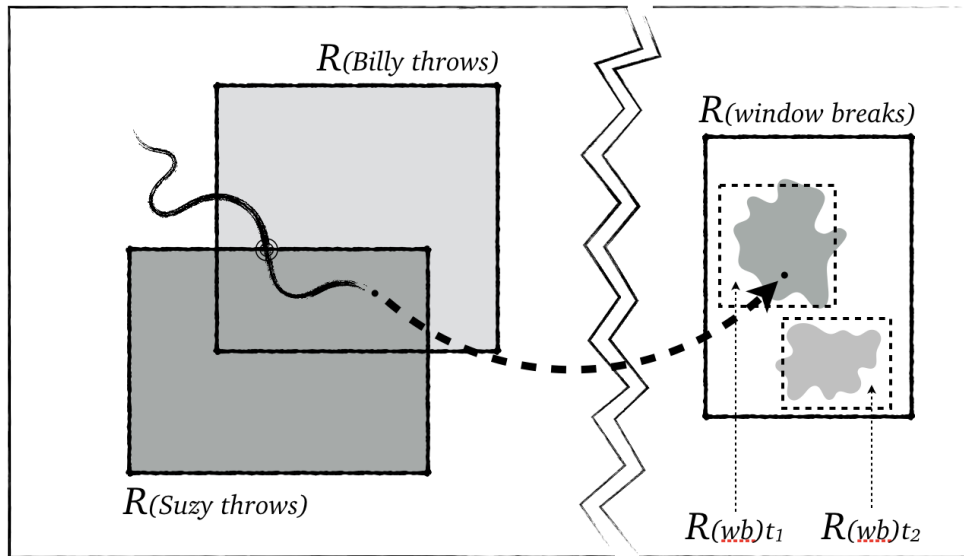

Figure 10.

R(wb)t₁: the window breaks at t_1; *R(wb)t₂*: the window breaks at t_2. *R(window breaks)* encompasses two distinct images of different descriptive states (see Fig. 4).

That is, late preemption cases capitalise on the fact that time evolution takes trajectories first into a descriptive state the physical states of which are related to a subregion of the general effect-region, which, under a more detailed description, turns out to be such that the actual trajectory never enters into it. This is why everyday causal intuition picks out Suzy's throwing as the cause of the window's breaking. Even if trajectories leading to the actual instance of the effect first enter into 'Billy throws but Suzy doesn't' which is a projective state of 'the window breaks', it is not a projective state of 'the window breaks at t_1' (but only of 'the window breaks at t_2'). The principal projective state of 'the window breaks at t_1' is 'Billy throws and Suzy throws', and the region the boundary of which the trajectories cross when they enter the principal projective state corresponds to Suzy's throwing.

Note that our solution here is similar to what (Lewis, 1986) proposes. However, whereas relying on a re-description of the effect seems to be an *ad hoc* move within

Lewis' framework, it is naturally implied by our framework as it recommends that the level of grain for the analysis of the cause and the effect should match each other (see Sec. 2.4). The proposed fine-graining of the effect here follows this guideline: it is governed by the level of grain dictated by the original description utilised at the cause side.

4.2 Absences and misconnection

So far, we have argued that our approach is quite effective in analysing cases that pose problems for counterfactual and INUS accounts of causation. Nevertheless, existing physical accounts of causation also perform quite good in this respect. So now we turn to another set of problem cases (absences and misconnection) that raises difficulties for current physical accounts of causation since they present scenarios in which causes and effects are not physically connected (Schaffer, 2000, 2004).

The particular case we focus on is *omission*, a variety of negative causation or causation by absences where an absence plays the role of the cause. (Here, again, we chose to discuss that variety that *prima facie* seems to pose the most serious problem for our approach. Accounting for prevention and double prevention follows straightforwardly and unproblematically from our framework.) Fig. 11 represents one of the classical examples of omission, the case where the gardener causes the plant's death by not watering it (while it also represents that the Queen might also be watering the plant; see Beebee, 2004). The specific scenario of Fig. 11 is such that the gardener did water the plant at some point in the past, but stopped doing it, which led to the death of the plant.

The principal projective state of the plant's death is the state when neither the gardener nor the Queen water the plant, and the trajectories enter this principal projective state through the boundary of the region corresponding to the property that the gardener is watering the plant. So 'the' cause of the plant's death is that this property ceases to be instantiated, i.e. that the gardener is not watering the plant (since as the trajectories cross the corresponding boundary they leave a state behind where the gardener is watering the plant and enter into a state where the gardener is not watering the plant). That is, our approach is in line with everyday causal intuition with regard to omission as well.
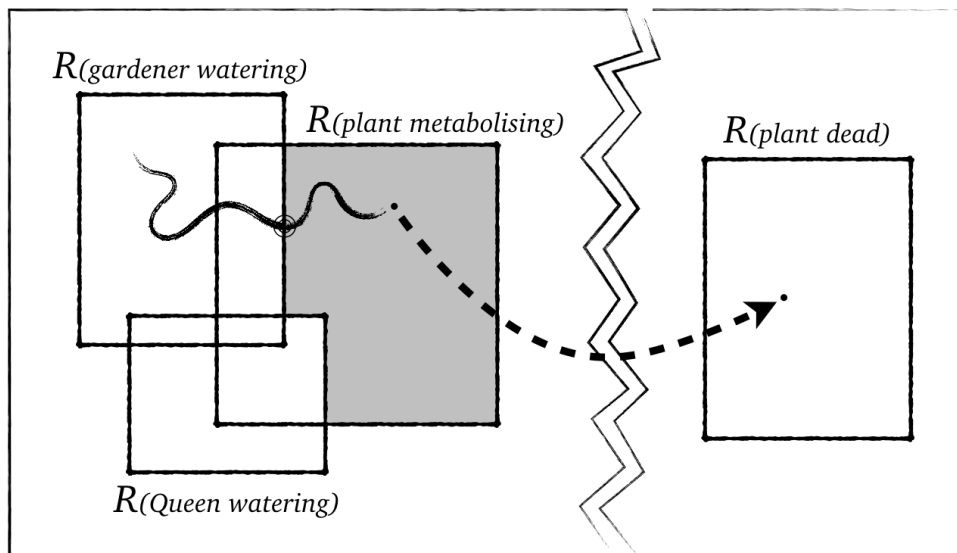
Figure 11.
The gardener stops watering the plant, the plant dies. The principal projective state of the 'plant dead' effect state is greyed out. As noted earlier, the double zig-zag line indicates that the effect region is not positioned correctly with regard to the properties characterising the cause. It is especially salient in this case: *R(plant dead)* is, in fact, the complementer region of *R(plant metabolising)*—exotic cases of keeping the plant alive put aside. This case is also a nice illustration of the importance of the 'within characteristic time bound' clause (see Fn. 8): all plants die within some period of time; the interesting thing from the perspective of this case is whether the actual plant dies within a time bound that is characteristic of death due to lack of water.

One might want to note, however, that our analysis above relies on a specific scenario, with a gardener who watered the plant in the past, but then stopped doing so. But what if it wasn't the gardener who has last watered the plant? Imagine that one day as the Queen wanders around she finds this plant, and gets mesmerised by its beauty so much that she decides to water it. The gardener had previously watered the plant many times, but then forgot about it and stopped watering it. Accidentally, it is the Queen who last waters the plant; after she leaves the plant stays there forlornly, and soon dies.

Fig. 12 represents this scenario. In this case, the trajectories enter the principal projective state from the state where the gardener is not, but the Queen is watering the plant, and the region the boundary of which the trajectories cross when they enter the principal projective state corresponds to the Queen's watering. But this conclusion (that the plant's death is caused by the Queen not watering it) seems to contradict with everyday causal intuition: the Queen's action was an unexpected and unprecedented event; it was the gardener who *should have* watered the plant.
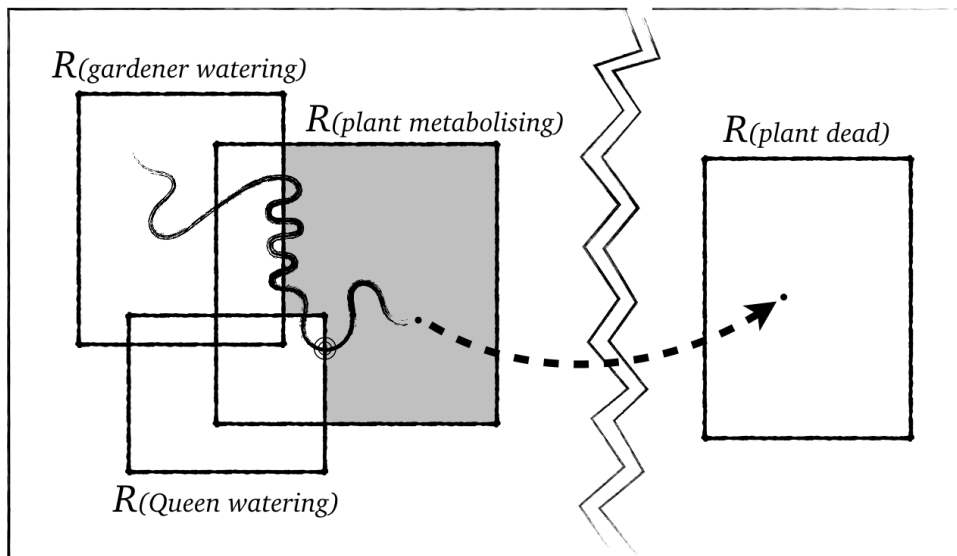
Figure 12.

The gardener stops watering the plant, the Queen waters it once, then the plant dies. In this case, the trajectory leading to the particular instance of the effect enters the principal projective state from the 'Queen watering' state. Actuality considerations pick out the Queen's not watering as the cause of the plant's death. However, typicality considerations pick out the gardener's negligence as the cause. See Sec. 3.1, and Fn. 16 & 17.

Causal intuitions about who *should have* watered the plant are based on typicality considerations.[17] We have seen earlier that within our framework it is possible to talk intelligibly about the typical cause of an effect. The typical principal projective state will be that projective state, which is crossed by the majority of trajectories leading to the physical states of the effect (see Sec. 3.1). If we combine this with how our framework solves the problem of selection (Sec. 3.4), then *'the' typical cause* of an effect is the instantiation of the property that the majority of the trajectories in question cross when they enter the typical principal projective state of the effect. So in our particular case, since the Queen's plants are—typically—taken care of by the gardener, it is the gardener's negligence that is the typical cause of the death of one of the Queen's plants.

To conclude this section, note two things. First, the way our framework accounts for omissions straightforwardly solves the 'too many negative causes' worry (Beebee, 2004; Lewis, 2004), since the descriptive states corresponding to possible alternative causes (i.e. the King's not watering, or Martians' not watering, etc.) will not be on the trajectories leading to the physical states of the actual instance of the effect. Second, our

---

[17] McGrath (2005) argues that causal intuitions in omission cases are governed by normative expectations. Normative expectations are grounded either in one's knowledge about what happens typically under some circumstances, or in morally or socially dictated obligations. The latter cases, however, fall outside what any theory of causation should rationally be expected to account for.

approach is able to account for cases of negative causation on the basis of the very same methodology that it uses to account for instances of positive causation, as absences (i.e. the non-instantiations of certain properties) correspond to regions of the state space very similarly to how occurrences (the instantiations of properties) do. In this respect, our framework clearly outperforms existing physical theories that need to evoke counterfactual considerations (and therefore resources not in their original toolkit—see Dowe, 2001, 2004).

5. Conclusion

Our framework approaches the task of accounting for causal claims from the perspective of dynamical systems. Causal systems are dynamical systems in the minimal sense that their physical states evolve with time. We propose that physical facts about such time evolution and the set of available linguistic descriptors are what ultimately determine the truth of causal claims.

The properties that are typically invoked by a causal discourse to describe certain states of affairs (i.e. to characterise causes and effects) correspond to regions in the state space of the causal system that include all those physical states that instantiate the properties in question. A descriptive state, which corresponds to a characterisation of a state-of-affair at the level of causal discourse, then, is an intersection of the regions corresponding to the properties utilised in the causal discourse to characterise the state-of-affair. A causal claim asserts a connection between two such descriptive states. These causal claims are reliable and robust, if there are real connections determined by time evolution between the two states in question such that a significant ratio of the physical states of the cause region evolve into the effect region—i.e. if the cause state is a projective state of the effect state. The causal claim is relevant, if the cause state is the principal projective state of the effect.

Besides being able to account for typicality, causally relevant factors, being 'the' cause, and cases of overdetermination, negative causation and misconnection, our approach is also able to reflect key commitments of existing accounts of causation. For example, due to the fact that trajectories (full descriptions of the system) play a foundational role in our approach our account is able to incorporate Mill's view (Mill, 1846),

and it also conforms well to John Norton's recent work on causation (Norton, 2003).[18] The fact that the principal projective state is crucial in determining the causally relevant factors of a given effect reflects the very same difference making intuition that is fundamental for difference making (Psillos, 2002) and interventionist (Woodward, 2003) accounts of causation. The fact that we determine projective states relative to how effects are defined by the causal discourse conforms our approach to contrastive accounts that see this feature of causation as crucial (Schaffer, 2005). Moreover, our focus on projective states in general can be seen as a Mackie-style move, since the non-accidental (relevant) properties defining a projective state form a set of INUS conditions (Mackie, 1974).

We see these similarities as virtues of our approach. They tell us that different existing accounts of causation focus on different features of causation, none of which in itself is able to fully characterise causal relations. Our approach, on the other hand, is integrative in the sense that it takes all of these insights seriously, incorporates them into one single account, and thus tries to overcome the classical dichotomy that thinks of causation either in terms of production or in terms of dependence (Hall, 2004). The integrative nature of our project could further be emphasised by demonstrating its compatibility with more pragmatic (e.g. interventionist) approaches to causation. Achieving this, i.e. developing links between our framework and the modelling literature is a specific aim of our future research.

Finally, note that the novel account of causation presented here has many possible fields of application that we have not even had a chance to mention. Consider, for instance, the problems of physicalism in general, or of mental causation in particular. The debates over whether physicalism is true or whether the mental can be causally efficacious centre around principles like the causal closure of the physical (Papineau, 2002) or the causal exclusion principle (Kim, 2005). However, as it has been argued by many in recent years, these principles disregard the growing consensus that the notion of causation only appears in macro-level descriptions—that at the micro-level one is only allowed to talk about determinism (see e.g. Price and Corry, 2007; Papineau, 2013). Our approach to causation is perfectly compatible with this consensus, and thus could provide new foundations for reframing these traditional debates.

---

[18] Even though more thorough investigation is required to assess the compatibility of our approach with indeterministic dynamics.

Acknowledgements

Funding

References

Andersen, H. (2017), "Patterns, Information, and Causation." *Journal of Philosophy*, 114, 11, 592-622.

Beebee, H. (2004), "Causing and nothingness." In *Causation and Counterfactuals* (J. Collins, N. Hall, and L. Paul, eds.), 291–308, Cambridge: MIT Press.

Dowe, P. (2000), *Physical Causation*. New York: Cambridge University Press, New York.

Dowe, P. (2001), "A counterfactual theory of prevention and 'causation' by omission." *Australasian Journal of Philosophy*, 79, 216–226.

Dowe, P. (2004), "Causes are physically connected to their effects: why preventers and omissions are not causes." In *Contemporary Debates in Philosophy of Science* (C. Hitchcock, ed.), 189–196, Blackwell.

Fair, D. (1979), "Causation and the flow of energy." *Erkenntnis*, 14, 219–250.

Hall, N. (2004), "Two concepts of causation." In *Causation and Counterfactuals* (J. Collins, N. Hall, and L.A. Paul, eds.), 225–276, MIT Press.

Hart, H. L. A. and A. M. Honore (1985), *Causation in the Law*. Oxford: Clarendon Press.

Hitchcock, C. (2012), "Theories of causation and the exclusion argument." *Journal of Consciousness Studies*, 19, 5-6, 40-56.

Kim, J. (2005), *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kistler, M. (2006), *Causation and Laws of Nature*. Oxford: Routledge.

List, C. and P. Menzies (2009), "Non-reductive physicalism and the limits of the exclusion principle." *Journal of Philosophy*, 106, 475-502.

List, C. and M. Pivato (2015), "Emergent Chance." *Philosophical Review*, 124, 1, 119-152.

List, C. and M. Pivato (forthcoming), "Dynamic and Stochastic Systems as a Framework for Metaphysics and the Philosophy of Science." *Synthese*, http://philsci-archive.pitt.edu/12371/

Lewis, D. (1986), "Causation." In *Philosophical Papers Volume II*, 159–213, Oxford: Oxford University Press.

Lewis, D. (2004), "Void and object." In *Causation and Counterfactuals* (Collins J., Hall N., and Paul L. A., eds.), 277–290, MIT Press.

Mackie, J. L. (1974), *The Cement of the Universe*. Oxford: Oxford University Press.

McGrath, S. (2005), "Causation by omission: A dilemma." *Philosophical Studies*, 123, 125–148.

Mill, J. S. (1846), *A System of Logic*. New York: Harper and Brothers.

Norton, J. D. (2003): "Causation as Folk Science." *Philosophers' Imprint*, 3, 4; reprinted in *Causation, Physics and the Constitution of Reality* (H. Price and R. Corry, eds.), 11-44, Oxford: Oxford University Press.

Papineau, D. (2002), *Thinking about Consciousness*. Oxford: Clarendon Press.

Papineau, D. (2013), "Causation is Macroscopic but not Irreducible." In *Mental Causation and Ontology* (Gibb, S., and Lowe, E, eds.), 126-152, Oxford: Oxford University Press.

Price, H., and R. Corry (2007), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Clarendon Press.

Psillos, S. (2002), *Causation and Explanation*. Chesham: Acumen.

Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Schaffer, J. (2000), "Causation by disconnection." *Philosophy of Science*, 67, 285–300.

Schaffer, J. (2004), "Causes need not be physically connected to their effects: The case for negative causation." In *Contemporary Debates in Philosophy of Science* (C. Hitchcock, ed.), 197–216, Oxford: Blackwell.

Schaffer, J. (2005), "Contrastive causation." *The Philosophical Review*, 114, 297–328.

Strevens, M. (2004), "The Causal and Unification Approaches to Explanation Unified – Causally." *Noûs*, 38, 1, 154-176.

Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Yablo, S. (1992), "Mental Causation." *The Philosophical Review*, 101, 245-280.