Erkenn (2019) 84:939–951
https://doi.org/10.1007/s10670-018-9988-5

Old Problems for the Agency Theory of Causal Discourse

Shyane Siriwardena[1]

Received: 10 July 2017 / Accepted: 16 February 2018 / Published online: 21 February 2018

Abstract Price's (Br J Philos Sci 42(2):157–176, 1991; 44(2):187–203, 1993 (with Peter Menzies); 2007, 2017) agency theory of causation has takes itself to provide a use-theory of our causal discourse. The theory's aim is to describe the rules implicit to our linguistic behaviour when we describe things in causal terms. According to this theory, the rules governing our use of the concept of causation are based on our perspective as agents and our associated experiences of manipulating events. I argue that the observed relation between agency and our concept of causation cannot exhaustively describe the conditions under which we enter into causal discourse. In particular, I demonstrate that the agency theory faces familiar problems with accounting for causal ascriptions to token cases.

Our language is littered with causal talk. We break things, move things, shape things. So much of the language we use to describe events characterises them as, in some respect, causing or being caused. Much of the philosophical discourse on causation concerns the metaphysics of the causal relation; but, Price's (1991, 1993 (with Peter Menzies), 2007, 2017) *agency theory of causation* has occupied itself with providing an account of our causal discourse. The theory's aim—or one of them—is to describe the rules implicit to our linguistic behaviour when we describe things in causal terms. According to this theory, the rules governing our use of the concept of causation are based on our perspective as agents and our associated experiences of manipulating events. In this paper, I will argue that this observed relation between agency and our concept of causation cannot exhaustively describe the conditions under which we enter into causal discourse. In particular, I will demonstrate that the agency theory faces some very old problems with accounting

✉ Shyane Siriwardena
s.siriwardena@leeds.ac.uk

[1] School of PRHS, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, UK

for causal ascriptions to token cases. I will argue that, even when modified, the agent theoretic use-rules for our concept of causation fall short of explicating all applications of our concept of causation to token events.

1 The Agency Theory, Old and New

1.1 Introducing the Agency Theory

It is widely recognised that not all regularities are causal. Much ink has been spilt on formalising this intuitive distinction, and the agency theory is best seen as part of this tradition. It begins with the idea most notably articulated in Cartwright (1979) that what distinguishes causal regularities from non-causal ones is that the former can be exploited by agents in order to achieve their ends. Causes, the thought goes, are effective strategies for bringing about their effects. Price's (1991, 1992, 1993 (with Peter Menzies), 2017) agency theory takes this idea one step further; on his account, we call a cause a cause *because* it is an effective strategy for achieving its effect. Price has defended versions of this theory in several different places, demonstrating that it can do much philosophical work, including: (a) explaining why agency and causation are related at all, (b) explaining why the arrows of temporal and causal asymmetry point in the same direction (1992, 2009) (c) avoiding cases of spurious causation that plague probabilistic accounts of causation (1991), and (d) providing an irenic solution to the Newcomb Problem in decision theory (1991, 2012). In short, the theory holds much promise; if successful, it could provide the answer to a number of different questions and problems in the philosophy of causation and beyond.

Broadly, Price argues that "the effects of an event A are those events to which A would provide a means" (1992: 261); in other words, "if in the context of means—end deliberation to realise A as the immediate product of a free action would be to raise the probability of B, then B is thought of as an effect of A" (1992: 261). More recently, Price has described his project as one concerned with the "*concept* of causation" and in particular with its "use", where by 'use' he means our "linguistic behaviour" (2017: 78, emphasis added).[1] However, Price has never given a precise statement of this version of the theory. Thus, I will construct an explicitly use-theoretic version of the agency theory on Price's behalf.

1.2 Use-Theoretic Entry Rules

As mentioned above, this paper will be concerned with a *use-theoretic* version of the agency theory; that is, a theory that is concerned with explicating the rules of use particular to causal discourse. In general, use-theories set out to describe the rules that best fit our linguistic behaviour with respect to the discourse in question. Some

[1] He also describes his project as involving a "genealogical" explanation of the origin and function of our concept. This part of his project provides motivation for the appeal to agency, but does not otherwise bear on the use-theory of our present concept. As such I will not discuss this part of Price's enterprise here.

of these rules are called "entry" rules (cf. Sellars 1954).[2] These are the rules that govern our 'entrance' into a discourse, where to enter a discourse is to transition from a non-linguistic state (e.g. a perceptual-state or belief-state) to a linguistic response (e.g. an assertion like '*x* is *F*'). To take a simple example, the entry rule for colour-discourse might tell us to utter or assent to 'this is red' when we have a red sense-experience.

Importantly, it need not be, and indeed often is not, the case that concept-users are explicitly aware of the rules for a discourse. The rules are said to be *implicit* to the discursive practice when the behaviour of participants in the practice conforms to those rules. A rule is implicitly followed when the concept-user is not aware of the rule governing the use of the concept in question, but their linguistic behaviour nevertheless exhibits a pattern consistent with following them.[3] And as theorists, it is from the behaviour of competent concept-users that we are meant to glean these rules of use.

1.3 The Use-Theoretic Agency Theory

In their 1993 paper, Menzies and Price define an "agent probability" as a conditional probability that is "assessed from an agent's perspective under the supposition that the antecedent condition is realised ab initio, as a free act of the agent concerned" (190). They explain that *A* is an effective means for achieving *B* "just in case [P(*B*|*A*)] is greater than [P(*B*|¬*A*)]" (190, changed to match notation), where these conditional probabilities are agent probabilities. Armed with these concepts, they argue that "*A* is a cause of a distinct event *B* just in case bringing about the occurrence of A would be an effective means by which a free agent could bring about the occurrence of *B*" (189). Since we are now interested in a version of this theory on which the explanandum is our causal linguistic practice, I will take it that claims of the sort "an event *A* is a cause of a distinct event *B*" in fact concern our use of causal concepts—that is, when we would *say* of *A* that it is a cause of *B*.

Translating the 1993 statement into a use-theory, we can construct the following agent-theoretic rule for the entry into causal discourse:

[2] There are other rules that pertain to the use of a discourse as well. Sellars, for instance, also describes *exit* rules (rules that govern transitions from a linguistic state to a non-linguistic state, such as an action). Discourses also involve *inference* rules (for moving between linguistic states). As this paper only concerns entry rules, I will not discuss these other rules any further.

[3] To be clear, the concept-user may or may not recognise or accept said rule if it were presented to them explicitly; but the rule may still be correct, even if the concept-user rejects it. It should also be noted that the fact that the rule is implicit does not preclude concept-users from correcting or instructing one another in their usage.

(C-Entry) The concept-user will assent to "*A* is a cause of *B*", if and only if their credences are such that $Cr(B|A) > Cr(B|\neg A)$, where the antecedent event is conceived of as being brought about by the concept-user.[4,5]

Where, to clarify, concept-users' credences can respect the described inequality even if they would not themselves describe their credences that way. If this were not the case, the rules would be far too demanding, since very few ostensibly competent users of the concept of causation conceive of their beliefs in credential terms.

The reader will note that while Menzies and Price used a general probability function in their account, I have stated the use-theoretic version of their theory in credential terms. This is in keeping with Price's understanding of effective strategies. Price (1991, 2012) defends a version of Evidential Decision Theory according to which an agent should regard *A* as an effective means of bringing about *B* just in case their credence in *B*'s occurrence given that they bring about *A* is greater than their credence in *B*'s occurrence given that they do *not* bring about *A*. Furthermore, Price (2012) explicitly applies this to his account of causation, arguing that "the information that events of type *A* are (positively) causally relevant to events of type *B* is the information that rationality requires that an agent contemplating an action of type *A* take it to be positively evidentially relevant to the occurrence of an outcome of type *B*" (2012: 514). Thus, stating the agency theory in credential terms remains in the Pricean spirit.

The use-theoretic agency theory now to hand, I will now present a series of cases, some of which will be familiar from the literature on counterfactual and probabilistic theories. However, despite their familiarity, these cases have not, before now, been deployed in this way. It is well known that the problems I will consider apply to theories that try to give an account of the extension of '*A* causes *B*'. What is less clear is that they apply to accounts of our linguistic behaviour. In the next section, I will show that, when modified to apply to a use-theory, these old problems show that the agency theory fails to adequately account for causal ascriptions to token cases.

2 Counterexamples

2.1 Simple Token Events

Let's begin with a bog-standard case involving token events from the past. Suzy, killing time on her own, is throwing stones at a bottle. Suzy throws a stone

[4] '*A*' and '*B*' are variables for events, whereas it is propositions that must take the argument place in probability functions. As such, in this paper, when any event variable *A* figures in any probability function, it will be elliptical for '*A* occurs'.

[5] In what follows, I will drop the 'where the antecedent event is conceived of as being brought about by the concept-user' locution for brevity. *Unless otherwise specified*, all conditional credences should be understood as agent probabilities, where the antecedent condition is conceived of as being brought about by the relevant concept-user.

(*Throw*$_S$),[6] it strikes the bottle, and the bottle shatters (*Shatter*). It is clear that *Throw*$_S$ is a cause of *Shatter*. But, (C-Entry) cannot accommodate this use of the concept. (C-Entry) requires that the concept-user's credences be such that $Cr(B|A) > Cr(B|\neg A)$. If the events in this inequality are meant to be *token* events, in the present bottle-breaking case, the observer must have credences such that $Cr(Shatter|Throw_S) > Cr(Shatter|\neg Throw_S)$, where the events conditioned upon (in this case, *Throw*$_S$ and $\neg$*Throw*$_S$) are conceived of as being brought about by the observer. But, we are imagining that the causal ascription in this case is made by an observer who has watched the scenario play out. In other words, the observer believes that those events have *already occurred*. So what could it mean for the observer to have beliefs such that *Shatter* would be more likely if they brought about an event that has already happened, rather than one that has not?

Since we have just stipulated that the observer witnessed the scene from start to finish, we know that they have just watched *Throw*$_S$ and *Shatter* occur. As a result, on a plausible understanding of token-level events, it is simply not possible for the observer to think that the event *Throw*$_S$ is (potentially) under their control. That very event cannot occur again—only an event of the *type* to which *Throw*$_S$ belongs could do so. Similarly for *Shatter*. The observer can no more have beliefs consistent with taking *Shatter* to be a possible outcome of their action than they can take *Throw*$_S$ to be in their control because it is part of our idea of a token event's being in the past that it has already occurred, and cannot happen again. And this is so, even in those cases where we have a less than certain credence that the events in question have occurred. In this case, the observer has credences near 1 in the occurrence of *Throw*$_S$ and *Shatter* because they witnessed the events' occurrences, but even when we have a relative low credence in the occurrence of a past event (e.g. the event of my climbing a mountain yesterday), we nevertheless do not think that that event is *now* in our direct or indirect control. If this is right, and the observer can neither conceive of past events as being brought about by them, nor conceive of them as the possible result of a free action, then their credences about such events cannot have the necessary properties to be agent probabilities. A fortiori, the observer's credences vis-à-vis *Throw*$_S$ and *Shatter* cannot respect the agent probabilities outlined in (C-Entry). Thus, we have a case in which the concept-user would assent to the claim '*Throw*$_S$ is a cause of *Shatter*', but their credences are not such that $Cr(Shatter|Throw_S) > Cr(Shatter|\neg Throw_S)$. This is a violation of (C-Entry).

One obvious move here would be to restate the agency theory counterfactually. We might alter the account to require that the following counterfactual be true of me in order for '*Throw*$_S$ is a cause of *Shatter*' to be assertable by me: at a time before the occurrence of *Throw*$_S$, my credence in the occurrence of *Shatter* given that I brought about *Throw*$_S$ would have been greater than my credence in *Shatter* given that I brought about $\neg$*Throw*$_S$. While a counterfactual solution might work in this case, such a theory would fail at the next hurdle—namely, when faced with preemption cases. I will demonstrate this in detail below, and so postpone any

[6] A brief note on the naming conventions I will follow in this paper. Token events will be identified by *unbolded, italicised names*; type-level events will be identified by ***bolded, italicised names***. When using variables, token events will be identified by lower-case letters (e.g. *a, b, c*), and type-level events will be identified by bolded upper-case letters (e.g. ***A, B, C***).

further discussion of a counterfactual solution until then. For now, I will assume the agency theorist must find an alternative solution.[7]

A plausible option is to modify (C-Entry) such that the agent-theoretic inequality involves credences about the *types* to which the relevant tokens—in this case *Shatter* and *Throw*$_S$—belong, on some canonical description. Modifying the statement of the agency theory accordingly, we can state the rule applicable to token causal ascriptions, as follows:

(C-Entry$_1$) When the question arises, a competent concept-user will assent to the claim '*a* is a cause of *b*' if and only if (a) the concept-user believes *a* and *b* both occurred; and (b) the concept-user's subjective probabilities are such that, given corresponding type-level events $\boldsymbol{A}$ and $\boldsymbol{B}$, $\mathrm{Cr}(\boldsymbol{B}|\boldsymbol{A}) > \mathrm{Cr}(\boldsymbol{B}|\neg\boldsymbol{A})$, where the antecedent events are taken to be brought about by the concept-user

Before testing this new rule, there are two details that require clarification. The first concerns the quantifiers ranging over $\boldsymbol{A}$ and $\boldsymbol{B}$. Earlier, I said that '*A*' and '*B*' are elliptical for '*A* occurs' and '*B* occurs'. But now that the variables concern type-level events, it is simply not clear what proposition is expressed by '$\boldsymbol{A}$ occurs', and so unclear what '$\mathrm{Cr}(\boldsymbol{B}|\boldsymbol{A}) > \mathrm{Cr}(\boldsymbol{B}|\neg\boldsymbol{A})$' expresses. Arguably, the latter is best seen as ambiguous between: (1) for all $\boldsymbol{A}$-events, there is some $\boldsymbol{B}$-event such that $\mathrm{Cr}(b|a) > \mathrm{Cr}(b|\neg a)$; (2) for all $\boldsymbol{B}$-events, there is some $\boldsymbol{A}$-event such that $\mathrm{Cr}(b|a) > \mathrm{Cr}(b|\neg a)$; and (3) for some subset of $\boldsymbol{A}$-events and $\boldsymbol{B}$-events, $\mathrm{Cr}(b|a) > \mathrm{Cr}(b|\neg a)$.

Neither (1) nor (2) can be appropriate for type-level credences that ground corresponding token-level causal claims. The former requires that *whenever* an $\boldsymbol{A}$-event occurs it increases my credence in the occurrence of some $\boldsymbol{B}$-event; but this is far from required for me to have the relevant token-level causal belief that some *a* is a cause of *b*. For instance, let's suppose I think that Jonny's smoking is a cause of his lung cancer. It is nevertheless not the case that *every* smoking event increases my credence in the occurrence of a cancer event; e.g. I do not think that cancer is any more likely in a smoker that already has cancer. In general then, if we read the conditional inequality in accordance with (1), we would almost never have the type-level beliefs called for for a given token-level claim.

Contrastingly, (2) requires that whenever a $\boldsymbol{B}$-event occurs, there is some $\boldsymbol{A}$-event such that the latter increased my credence in the occurrence of the former. But neither is this necessary for every belief of the form *a* is a cause of *b*. Return to my belief that Jonny's smoking is a cause of his cancer; I believe this, but also believe that some cancer patients have never smoked, so there are some $\boldsymbol{B}$-events such that there is no $\boldsymbol{A}$-event that increased my credence in its occurrence. Therefore, (2) is similarly over-demanding. In general, the problem is that where causal beliefs are

[7] Readers familiar with Price might object that I have neglected his solutions to the problem of unmanipulable causes, of which my case is arguably an example. Menzies and Price (1993) suggest we should reason by analogy from manipulable cases, and Price (2017) suggests we use "extension principles" (2017: 91). But neither of these will be effective as such solutions would still be vulnerable to counterexamples arising from preemption cases and probability-lowering causes. (My thanks to an anonymous referee for raising this point).

concerned, it is simply not the case that we take the relation to hold *universally*. Thus, for the remainder of the paper, I will take it that some version of (3) is appropriate; i.e. that $Cr(\boldsymbol{B}|\boldsymbol{A}) > Cr(\boldsymbol{B}|\neg\boldsymbol{A})$ should be read as expressing the quantified claim, 'for some significant subset of $\boldsymbol{A}$-events and $\boldsymbol{B}$-events, $Cr(b|a) > Cr(b|\neg a)$'. I will not, however, try to define or quantify a 'significant subset'. For the purposes of what follows, I will rely on an intuitive understanding of this notion.

The second detail concerns negative events. Notice that in (C-Entry$_1$)(b), the conditional probability states that the concept-user's credences must be such that their credence in $\boldsymbol{B}$ given $\boldsymbol{A}$ is greater than that in $\boldsymbol{B}$ given $\neg\boldsymbol{A}$. However, if $\neg\boldsymbol{A}$-type events can simply be any non-$\boldsymbol{A}$ event, there will be significant subsets of $\neg\boldsymbol{A}$-events for which this conditional credence does not hold. Suppose, for instance, that we are concerned with my causal beliefs about window-breakings and ball-throwings. I certainly believe that, in general, throwing balls at windows with sufficient force causes windows to break. According to the agency theory, it follows that my credences must be such that $Cr(\boldsymbol{Breaking}|\boldsymbol{Throwing}) > Cr(\boldsymbol{Breaking}|\neg\boldsymbol{Throwing})$. However, if the $\neg\boldsymbol{Throwing}$ events I'm considering are ***Gunshot*** events, then my credences will not respect this inequality; indeed, they will very likely be such that $Cr(\boldsymbol{Breaking}|\boldsymbol{Throwing}) < Cr(\boldsymbol{Breaking}|\neg\boldsymbol{Throwing})$.

Of course, intuitively, when we consider $\neg\boldsymbol{Throwing}$ events, we don't mean to include ***Gunshot*** events under that heading. Instead, we tend to have in mind the bare non-occurrence of a throwing event. For instance, a non-throwing event is currently occurring at my desk, but not in virtue of the occurrence of an event non-compossible with a throwing event. Rather, such an event is occurring in virtue of the fact that the set of positive events occurring at my desk does not include a throwing event. Applying this thought to the agent-theoretic inequalities, $Cr(\boldsymbol{B}|\boldsymbol{A}) > Cr(\boldsymbol{B}|\neg\boldsymbol{A})$ should be understood as pertaining to our credences about the occurrence of $\boldsymbol{B}$-events when we *add* the occurrence of an $\boldsymbol{A}$-event to some set of circumstances, as compared to when do not add such an event. To formalise this, take the set of positive event-types (where each event is identified under some canonical description) occurring just prior to $\boldsymbol{A}$ in the relevant circumstances.[8] Call this set $\boldsymbol{S}$. With $\boldsymbol{S}$, we can more accurately state the agent-theoretic inequality as follows: $Cr(\boldsymbol{B}|\boldsymbol{A}\cdot\boldsymbol{S}) > Cr(\boldsymbol{B}|\boldsymbol{S})$. The result of this is that claims about the relationship between bringing about some $\boldsymbol{A}$ and our credence in the occurrence of some $\boldsymbol{B}$ will always be relativised to some set of circumstances. In what follows, I will take it that we are always conditionalising on some relevant set of circumstances, so for simplicity I will drop the '$\boldsymbol{S}$' from the inequality. Thus, we will say, for instance, that my credences must be such that $Cr(\textbf{Breaking}|\textbf{Throwing}) > Cr(\textbf{Breaking})$ in the bottle-breaking case. In general, then, we can amend (C-Entry$_1$) to read as follows:

[8] These can be past, present, future, or hypothetical, since we make causal claims about events in all such circumstances.

(C-Entry$_1$) When the question arises, a competent concept-user will assent to the claim '*a* is a cause of *b*' if and only if (a) the concept-user believes *a* and *b* both occurred; and (b) the concept-user's subjective probabilities are such that, given corresponding type-level events ***A*** and ***B***, Cr(***B***|***A***) > Cr(***B***), where the antecedent events are taken to be brought about by the concept-user

2.2 Preemption Cases

To test (C-Entry$_1$), consider the same stone-throwing scenario as above, but this time suppose that Billy has joined his friend for their usual game. Now Billy and Suzy—both expert rock marksmen—throw their rocks at a glass bottle at the same time, each with enough force to smash the bottle. Suzy, however, throws just a bit harder than Billy does, so her rock hits the bottle first and smashes it; Billy's rock then flies through the shards of glass where the bottle was just a split-second earlier. Here, I take it that any ordinary user of the concept of causation would identify Suzy's throw (*Throw$_S$*), not Billy's (*Throw$_B$*), as a cause of the bottle's shattering. Further, I take it that there is no plausible case to be made for the claim that this causal ascription is mistaken. It is simply incredible to think that any theory that takes Billy's throw in this case to be rightfully called a cause, or Suzy's throw not to be, could count as a theory of our ordinary concept of causation.

Before I proceed, I return to the promise made in 2.1. I said then that this case also threatens the counterfactual formulation of the agency theory. Suppose that (C-Entry) were expressed as follows: Token event *a* is properly called a cause of a distinct token event *b* just in case (a) the concept-user believes *a* and *b* both occurred, and (b) the concept-user's subjective probabilities at a time prior to the occurrence of *a* would have been such that Cr(*b*|*a*) > Cr(*b*|¬*a*), where the antecedent event is taken to be brought about by the concept-user. In the case just described, it is false that the concept-user's credences would have been such that Cr(*Shatter*|*Throw$_S$*) > Cr(*Shatter*|¬*Throw$_S$*) at a time just prior to *Throw$_S$* since they would know that *Throw$_B$* was going to occur. As such, their credence in *Shatter* would not be any greater given *Throw$_S$*, but they would assent to the claim that *Throw$_S$* is a cause of *Shatter*, violating the counterfactual version of (C-Entry). Further, it is no good arguing that *Throw$_S$* would marginally increase the concept-user's credence in *Shatter*, since the same would be true of *Throw$_B$*. In this case (C-Entry) would have the observer count *Throw$_B$* as well as *Throw$_S$* a cause of *Shatter*. A clearly unacceptable result.

So much for the counterfactual strategy. Turning back to (C-Entry$_1$), the observer in this case satisfies the first condition: they believe *Throw$_S$* and *Shatter* to have occurred. Now suppose the observer also satisfies the second condition; that is, suppose their credences are such that Cr(***Shatter***|***Throw$_x$***) > Cr(***Shatter***|¬ ***Throw$_x$***), where ***Throw$_x$*** is taken to be brought about by the concept-user. In virtue of satisfying conditions in (C-Entry$_1$) relative to *Throw$_S$*, it follows by that rule that they will assent to the causal claim '*Throw$_S$* is a cause of *Shatter*' when the question arises. This is as we expected. Unfortunately, the observer also satisfies the two

conditions in (C-Entry$_1$) relative to *Throw*$_B$. The observer believes *Throw*$_B$ and *Shatter* to have occurred (for they witnessed both events occur), and, as was just stipulated, their credences are such that Cr(***Shatter***|***Throw***$_x$) > Cr(***Shatter***|¬***Throw***$_x$). Since *Throw*$_B$ is clearly a token of ***Throw***$_x$, (C-Entry$_1$) demands that the observer assent to the claim '*Throw*$_B$ is a cause of *Shatter*'. But, no competent concept—user would do this. Thus, we have a violation of (C-Entry$_1$) in virtue of the relevant conditions for assertion licensing an intuitively unacceptable claim.

Perhaps the problem was not with (C-Entry$_1$), but with the type-level description we have taken to be appropriate here. For example, *Throw*$_S$ and not *Throw*$_B$ falls under the event-type that involves not just the throwing of a stone at a bottle, but also the stone's impacting the surface of the bottle. Let's call this event-type '***Throw***$_x$***-Impact***'. Now, supposing the observer's credences are such that Cr(***Shatter*| *Throw***$_x$***-Impact***) > Cr(***Shatter***|¬ ***Throw***$_x$***-Impact***), only *Throw*$_S$ will be properly described as a cause of *Shatter* by (C-Entry$_1$). This looks promising. But, a problem arises for anyone whose credences are such that both

(a) Cr(***Shatter*| *Throw***$_x$) > Cr(***Shatter***|¬ ***Throw***$_x$) and,
(b) Cr(***Shatter*| *Throw***$_x$***-Impact***) > Cr(***Shatter***|¬ ***Throw***$_x$***-Impact***).

If that person makes causal claims based on both (a) and on (b)—as (C-Entry$_1$) seems to require—then they ought to call both *Throw*$_B$ and *Throw*$_S$ causes of *Shatter*. And there does not seem to be a principled way of identifying a unique, correct type-level event for any given token event. Indeed, it is hard to see how to do so in way that isn't viciously circular. For instance, in this case, ***Throw***$_x$***-Impact*** cannot just be the event of a throw that is *followed by* the event of that stone's impact. Suppose some half an hour after Billy threw his stone someone else picks up that very stone and shatters a bottle with it; in this case, Billy's throw would fall under the type ***Throw***$_x$***-Impact***, but it is surely wrong to say that Billy's throw caused the later shattering of a bottle. Indeed, it is not clear how to appropriately specify the events of the type ***Throw***$_x$***-Impact*** without referring back to the causal relation in question—i.e. events where a stone is thrown and that throw causes said stone to impact the bottle.

An alternative response to preemption cases that avoids the appeal to different type-level descriptions involves an appeal to causal chains. Instead of incorporating the impact of Suzy's stone on the bottle into the description of the antecedent throwing event, it can be included as a distinct event *Contact*$_S$ (i.e. the event of Suzy's stone contacting the surface of the bottle) that was caused by *Throw*$_S$ (and not by *Throw*$_B$) and that caused *Shatter*. The idea is that the bottle-shattering case is resolved as follows. The observer believes each of *Throw*$_S$, *Throw*$_B$, *Contact*$_S$, and *Shatter* to have occurred. Further, they have credences such that Cr(***Contact***$_x$|***Throw***$_x$) > Cr(***Contact***$_x$|¬***Throw***$_x$) (in natural language, for all X, and for some significant subset of ***Contact***$_x$-events and of ***Throw***$_x$-events, the observer's credence in occurrence of X's stone contacting the bottle given that they bring about X's throwing a stone is greater than their credence in the occurrence X's stone contacting the bottle); they also have credences such that Cr(***Shatter***|***Contact***$_x$) > Cr(***Shatter***|¬***Contact***$_x$) (in

natural language, for all X, and for some significant subset of $\boldsymbol{Contact_x}$-events and of ***Shatter***-events, the observer's credence in the occurrence of the bottle-shattering given that they bring about the event of X's stone making contact with the bottle is greater than their credence in the occurrence of the bottle-shattering). It follows, that if the question arises, the observer should assent to '$Throw_S$ is a cause of *Shatter*'. Moreover, since the observer does not believe an event $Contact_B$ to have occurred, the same cannot be said of $Throw_B$. Therefore, on this kind of rule, $Throw_S$ and not $Throw_B$ is a cause of *Shatter*.

Such a strategy faces two challenges. First, to specify the new entry rule in a way that does not entail (C-Entry$_1$) but still preserves the possibility of making causal ascriptions to cases where the concept-user does not have beliefs about intervening events. Second, to avoid familiar counterexamples to transitivity. But even if we supposed for the sake of argument that such a solution could be found,[9] the revised version of the theory would face counterexamples from probability-lowering causes.

2.3 Probability-Lowering Causes

Consider a case like the following [modified from Eells and Sober (1983)]:[10] you are teeing off in a round of golf. You swing and your ball rolls onto the green and is rolling toward the hole when a squirrel runs onto the course and kicks the moving ball in a new direction. Improbably, because of the nature of the green, the ball curves around and falls into the hole for a hole-in-one. Any competent concept-user would say that the squirrel's kick was a cause of the hole-in-one. So, the antecedent of the 'only if' direction of the biconditional is satisfied with respect to (C-Entry$_1$); but condition (b) in the consequent is not. I—and most others, I suspect—simply do not have credences such that

$$\mathrm{Cr}(\boldsymbol{Hole\text{-}in\text{-}one}|\boldsymbol{Squirrel\text{-}Kick}) > \mathrm{Cr}(\boldsymbol{Hole\text{-}in\text{-}one}|\neg\boldsymbol{Squirrel\text{-}Kick}).$$

Quite the opposite, in fact. I think bringing it about that a squirrel kicks my golf ball is a very good way to *prevent* a hole-in-one; i.e.

$$\mathrm{Cr}(\boldsymbol{Hole\text{-}in\text{-}one}|\boldsymbol{Squirrel\text{-}Kick}) < \mathrm{Cr}(\boldsymbol{Hole\text{-}in\text{-}one}|\neg\boldsymbol{Squirrel\text{-}Kick}).$$

Thus, (C-Entry$_1$) fails. And so too would any version of the rule including a credence-raising requirement.

You might think that, our credences would be different given some more precise description of the putative cause (e.g. squirrel-kick of force n at angle θ) and some more precise description of the context $\boldsymbol{S}$ (e.g. when the grass is dry, the ball is a certain distance from the hole, the slope is at a particular incline, etc.). However, if this proposal is to generalise to all cases, the agency theorist will require a clear method for identifying "appropriate" descriptions of events—a method that does not (tacitly or otherwise) appeal to the causal relation between $\boldsymbol{A}$ and $\boldsymbol{B}$. As such,

[9] Kvart's (2004) notion of an interfering factor could be used to provide such a solution.

[10] The first probability case involving a mediocre golfer is cited in Suppes (1970), but attributed to Deborah Rosen. See also Rosen (1978).

even if an appropriate method can be found for specifying ***S*** in the squirrel case, it is at best unlikely that the agency theory will be able to provide a principled rule for accommodating probability-lowering causal claims in this way.

It is worth noting that this problem concerning the appropriate description of events is not unique to the agency theory. It exists for any theory that appeals to type-level descriptions to make sense of token-level claims. That said, the problem is particularly pressing for the agency theory since it lacks a viable alternative for describing the rules implicit to making causal claims about token events.

2.4 A Note on Future Events

One might think that the agency theory is only vulnerable to the counterexamples above because the events concerned are in the past. After all, most of our reasoning about effective strategies concerns *future* events. But, as it turns out, the events' positions in time do not affect the strength of the counterexample.

Note that, where future token events are concerned, entry rule need not involve credences about event-types, for we can—and do—take such events to be in our (direct or indirect) control. Thus, I can deliberate, for instance, over whether the particular striking of a match I'm about to perform will cause the match's lighting. If my credences are such that $\mathrm{Cr}(\mathit{Light}|\mathit{Strike}) > \mathrm{Cr}(\mathit{Light}|\neg\mathit{Strike})$, then if asked, I should say of my future striking of a match that it will be causally relevant to the match's lighting. Indeed, the reverse seems to work as well. Suppose I tell you that my use of defoliant will cause the weeds to die; if I then did not regard my use of defoliant as making it more likely that the weeds die, we would think that I had misused or misunderstood my original causal claim. Given these reflections, we might state the agency theory's entry rule for future token events as follows:

($\text{C-Entry}_{\text{Future}}$) When the question arises, the concept-user will assent to "a is causally relevant to b" if and only if their credences are such that $\mathrm{Cr}(b|a) > \mathrm{Cr}(b|\neg a)$

But, the counterexamples above can very easily be modified to apply to this version of (C-Entry) as well.

For instance, consider the bottle-smashing case again, but this time imagine that we—the observers—are considering the situation at a time prior to either of the throws. To prevent intuitions concerning what we can or cannot know about other people's intentions from muddying matters, let's suppose that Billy and Suzy have been replaced by robots. RoboSuzy and RoboBilly are each programmed with excellent aim, but RoboSuzy is programmed to throw with more force than RoboBilly. Both robots have been programmed to throw their respective stones at exactly the same time. In this case, it seems right to assert that RoboSuzy's throw will cause the bottle to shatter. Now, according to ($\text{C-Entry}_{\text{Future}}$), I should think that RoboSuzy's throwing its stone is an effective strategy for bringing about a bottle-shattering; i.e. $\mathrm{Cr}(\mathit{Shattering}|\mathit{RoboSuzy's\ Throw}) > \mathrm{Cr}(\mathit{Shattering}|\neg\mathit{RoboSuzy's\ Throw})$. But my credences make this inequality false. My credences are instead such that $\mathrm{Cr}(\mathit{Shattering}|\mathit{RoboSuzy's\ Throw}) = \mathrm{Cr}(\mathit{Shattering}|\neg\mathit{RoboSuzy's\ Throw})$ because I know that RoboBilly is programmed to throw its stone as well. So, my

beliefs are such that I accept the causal claim without having the appropriate credences as outlined by (C-Entry$_{\text{Future}}$).

In general, the preemption problem will stand whenever the concept-user in question already has a credence close to 1 in the occurrence of the events in question, regardless of whether this is so because they have witnessed the events occur (as in the cases in the past), or because of other information they have (as in the cases in the future). The lesson to be learned here is this: what is relevant to the strength of the counterexample is not the events' position in time relative to that of the concept-user's ascriptions, but instead the concept-user's degree of belief in the occurrence of those events, regardless of when they did or will occur. Given this, modulo some modest changes, the arguments in the foregoing can also be applied to cases involving events in the future.

3 A Problem for Everyone

At this point, the reader might wonder why these much-discussed recalcitrant cases should be any more of a problem for the agency theorist than for anyone else. After all, aren't these cases a problem for everyone?[11] The short answer is, *yes*. Indeed, this is the very point that I wished to make. These cases are a problem for *everyone*—including the use-theoretic agency theory. The cases discussed here have long been recognised as problems for metaphysicians of causation, and those who would give analyses of causation. I have shown that changing the explanandum from causation itself, or the extension of the concept, to our causal linguistic practice does not get the agency theory off the hook. The counterexamples can be reformulated, as I have done above, to target a theory of our linguistic behaviour; thus, the foregoing argument provides a novel application of these old problems. The agency theory can't ignore these cases any more than Lewis's counterfactual theory could; and this is true despite the fact that they are engaged in two very different explanatory projects.

4 Conclusion

I have shown that the use-theoretic agency theory cannot provide an exhaustive account of our causal discourse. The reason for this is that *effective strategies* alone do not suffice for an account of our use of causal language in token cases. It is simply not the case that we only describe as causal those events that we think (or would think) would provide an effective means of bringing about a putative effect.

But we needn't give up the idea that our causal ascriptions are in some way related to our capacities as agents. For instance, Woodward's (2003) interventionism successfully avoids counterexamples of the kind discussed above. Thus, if the interventionist machinery can be put to use in a use-theoretic account of our concept

[11] Thanks to an anonymous reviewer for raising this important point.

of causation, there may yet be a proverbial port available to those sympathetic to the intuition that agency and our causal discourse are importantly related.

Acknowledgements I am grateful to the following colleagues for very helpful feedback on numerous versions of this paper: Arif Ahmed, Adam Bales, John Divers, Max Jones, Gail Leckie, Huw Price, Mathieu Rees, audience members at the Moral Sciences Club, and anonymous referees for this journal.

References

Cartwright, N. (1979). Causal aws and effective etrategies. *Noûs*, *13*(4), 419–437.
Eells, E., & Sober, E. (1983). Probabilistic causality and the question of transitivity. *Philosophy of Science, 50*(1), 35–57.
Kvart, I. (2004). Probabilistic cause, edge conditions, late preemption, and discrete cases. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world* (pp. 163–188). London: Routledge.
Menzies, P., & Price, H. (1993). Causation as a secondary quality. *The British Journal for the Philosophy of Science, 44*(2), 187–203.
Price, H. (1991). Agency and probabilistic causality. *The British Journal for the Philosophy of Science, 42*(2), 157–176.
Price, H. (1992). The direction of causation: Ramsey's ultimate contingency. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (pp. 253–267).
Price, H. (2007). Causal perspectivalism. In H. Price & R. Corry (Eds.) *Causation, physics, and the constitution of reality: Russell's republic revisited* (pp. 250–292). Oxford: OUP.
Price, H. (2012). Causation, chance, and the rational significance of supernatural evidence. *Philosophical Review, 121*(4), 483–538.
Price, H. (2017). Causation, intervention and agency. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a difference: Essays on the philosophy of causation* (pp. 73–98). Oxford: OUP.
Price, H., & Weslake, B. (2009). The time-asymmetry of causation. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 414–446). Oxford: OUP.
Rosen, D. (1978). In defense of a probabilistic theory of causality. *Philosophy of Science*, *45*(4), 604-613.
Schaffer, J. (2004). Counterfactuals, causal independence and conceptual circularity. *Analysis, 64*(284), 299–308.
Sellars, W. (1954). Some reflections on language games. *Philosophy of Science, 21*(3), 204–228.
Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: OUP.