A Critique of Yoav Shoham's Theory of Causal Reasoning

Antony Galton
Department of Computer Science
University of Exeter
Exeter EX4 4PT, U.K.
antony@uk.ac.exeter.dcs

Abstract

Causal reasoning is an essential part of a number of tasks that have been central to many endeavours in AI—notably planning and prediction, diagnosis and explanation. Recently it has become an object of study in its own right, drawing inspiration from the work of philosophers and logicians as well as more immediately AI-oriented concerns. In this paper I shall examine just one approach to causal reasoning, that advocated by Yoav Shoham in a recent book and article. In particular, I shall try to lay bare a number of assumptions underlying Shoham's work, all of which I shall call into question. Key assumptions are that causality is an epistemic notion, that causal reasoning is inherently non-monotonic, and that epistemic reasoning should be handled by means of modal logic. While arguing against these assumptions, I do not offer a specific causal theory of my own, but shall conclude with some suggestions as to the general lines which I feel such a theory ought to follow.

Introduction

Causal reasoning is an essential part of a number of tasks that have been central to many endeavours in AI—notably planning and prediction, diagnosis and explanation. Recently it has become an object of study in its own right, drawing inspiration from the work of philosophers and logicians as well as more immediately AI-oriented concerns.

Amongst recent attempts to provide a formal basis for causal reasoning in AI may be mentioned Lifschitz's use of circumscription to secure the desired non-monotonicity of causal inference (Lifschitz 1987), Pearl's interesting attempt to draw a formal distinction between two different kinds of default rule involved in causal reasoning, which he calls *causal* and *evidential* (Pearl 1988), and Geffner's use of a 'causal operator', C, which may be read, roughly, as 'we have a causal explanation for . . . ' (Geffner 1990).

In this paper I shall not examine these approaches in detail, but shall focus my discussion on the approach to causal reasoning advocated by Yoav Shoham in a recent book and article (Shoham 1988, Shoham 1990). In particular, I shall try to lay bare a number of assumptions underlying Shoham's work, all of which I shall call into question. I do not offer a specific theory of my own, but shall conclude with some suggestions as to the general lines which I feel such a theory ought to follow.

Shoham's Theory

A *causal theory*, for Shoham, is a collection of *causal statements*, which are rules of the form

$$\Box\phi_1 \wedge \Box\phi_2 \wedge \cdots \wedge \Box\phi_m \wedge \Diamond\chi_1 \wedge \Diamond\chi_2 \wedge \cdots \wedge \Diamond\chi_n \rightarrow \Box\psi$$

where the ϕ_i, χ_j, and ψ are atomic sentences; in addition there are some constraints on the time-reference of these atomic constituents (designed to ensure that an effect cannot precede its cause), but the details of these do not immediately concern us here.

The idea is that the ϕ_i express the active causes, while the χ_j express the background conditions (the 'causal field') which have to obtain in order for the active causes to have their effect. The modal operators are there in order to allow us to make default assumptions to the effect that unless we are explicitly told otherwise, we can take the background conditions to be present.

Shoham illustrates his ideas using an example concerning starting the engine of a motor-car. For the sake of variety, I shall use a somewhat different, though still essentially similar, example: pressing the button of the doorbell causes the doorbell to ring, so long as the circuit is connected up properly, the battery is not dead, the bell is not broken, and so on. In Shoham's system, we write [1]

$$\Box\text{Press-button}(t) \wedge \Diamond\text{All-working}(t) \rightarrow \Box\text{Bell-rings(t+1)}$$
$$\Box\text{Battery-dead}(t) \rightarrow \Box\neg\text{All-working}(t)$$
$$\Box\text{Wire-disconnected}(t) \rightarrow \Box\neg\text{All-working}(t)$$
$$\Box\text{Bell-broken}(t) \rightarrow \Box\neg\text{All-working}(t)$$

Shoham uses a non-monotonic inference mechanism called *chronological ignorance*, which in effect requires one to put off making positive assertions for as long as possible. By a 'positive assertion' is meant a statement of the form $\Box\phi$. In the example above, suppose we are given the statement

$$\Box\text{Press-button}(0).$$

Then one possibility is to assume, say, $\Box$Battery-dead(0), which would entail $\Box\neg$All-working(0), and hence would not allow us to conclude anything about time 1 (since the antecedent of the first conditional is falsified). Alternatively, we could refrain from making any positive assumptions about time 0, so that we do *not* have $\Box\neg$All-working(0); not having this is tantamount to having $\Diamond$All-working(0), and we can now use the first conditional to conclude that $\Box$Bell-rings(1).

Of the two alternatives outlined above, the principle of chronological ignorance prefers the second, since it puts off making positive assertions as late as possible—the first alternative makes a positive assertion about time 0 (namely $\Box$Battery-dead(0)) whereas the second only makes a positive assertion about time 1 (namely $\Box$Bell-rings(1)).

What do the modal operators mean?

I shall begin my critique of Shoham's theory by asking what, precisely, the modal operators $\Box$ and $\Diamond$ are supposed to mean. Shoham is rather vague about this; he says 'I will feel free to describe the $\Box$ modality as knowledge and belief interchangeably'. Thus we might choose to read Shoham's $\Box$ as something like 'It is known that . . . ', or alternatively as something like 'It is believed that . . . '. Or maybe the knowledge/belief should be located in a knowing/believing subject, say 'I know that . . . ' or 'I believe that . . . '. All this Shoham deliberately leaves inexplicit.

Consider now a causal rule such as

$$\Box\text{Press-button}(t) \wedge \Diamond\text{All-working}(t) \rightarrow \Box\text{Bell-rings(t+1)}.$$

If we read $\Box$ as 'I know that . . . ', then this rule must be read as saying something like

[1] Note that I have taken a few liberties with Shoham's notation; in particular I have 'unreified' his atomic propositions, writing for example Press-button(t) where Shoham would write True$(t, t,$Press-button)—see (Galton 1991).

(1) If I know that the button is pressed at t, and I do not know that all is not working at t, then I know that the doorbell rings at $t + 1$.

If on the other hand we read $\Box$ as 'I believe that . . . ', then the rule comes out rather as saying that

(2) If I believe that the button is pressed at t, and I do not believe that all is not working at t, then I believe that the doorbell rings at $t + 1$.

Now (1) will clearly not do. Suppose I know that the button is pressed at t; suppose also that, unknown to me, the battery is dead at t, so that it is not the case that all is working at t. In that case the bell will *not* ring at $t + 1$, and therefore, contrary to reading (1) of Shoham's rule, I cannot possibly know that it will ring then (one cannot know what is in fact false).

What about (2)? This is more plausible, because it is possible to believe a falsehood. But as it stands, it still can't be right: what we want is a *causal law*, not a psychological one, yet all (2) gives us is a statement of the dynamics of belief.

The causal law cannot in itself induce belief; at best it can give us a *reason* for believing that such-and-such an effect will follow, given certain causes. This suggests that the correct reading for the modal operator $\Box$ is something like 'There is reason to believe that . . . ' or, with a personal subject, 'I have reason to believe that . . . '. With this reading, our example comes out as

(3) If I have reason to believe that the button is pressed at t, and I do not have reason to believe that all is not working at t, then I have reason to believe that the doorbell rings at $t + 1$.

It is worth comparing this formulation with Geffner's account using the causal operator C (Geffner 1990). In Geffner's system, the rule about the door-bell might come out as

$$\text{Press-button}(t) \wedge \neg\text{Out-of-order}(t) \rightarrow C\text{Bell-rings}(t + 1)$$

i.e., roughly, 'if the button is pressed at t, and the bell is not out of order at t, then we can explain why the bell rings at $t + 1$'. (Note that, reasonably enough, 'we can explain why p' implies p in Geffner's system.)

There is clearly an affinity between the expressions like 'I have reason to believe that . . . ' and those like 'We can explain why . . . ', which appear in my renderings of Shoham's and Geffner's causal rules, but they are by no means identical in meaning; presumably a sufficiently rich causal theory might encompass both types of expression and articulate the relationship between them.

Does causal reasoning need modal operators?

Is 'I have reason to believe that . . . ' a modal operator? By a modal operator I mean an expression whose syntactic and semantic behaviour suits it to be represented by the $\Box$ of modal logic. Now the modal operator $\Box$ has the following salient properties

(a) It acts on propositions to form propositions: if ϕ is a proposition, then so is $\Box\phi$;

(b) It is significantly iterable: expressions such as $\Box\Box\phi$ make sense;

(c) It does not commute with negation: $\Box\neg\phi$ is, in general, not equivalent to $\neg\Box\phi$.

How does the expression ‘I have reason to believe that . . . ’ measure up to these properties?

As regards item (a), it may seem obvious that if ϕ is a proposition then so is ‘I have reason to believe that ϕ’—but of course this depends on what exactly one takes a proposition to be. Specifically, do we require a causal reasoning mechanism to be able to handle expressions of both these forms, and if so, should these expressions belong to the same syntactic category?

We may distinguish assertions about the domain (e.g., about door-bells and batteries) on the one hand from statements about our state of knowledge of the domain on the other. These two kinds of statements belong at different levels, for example ‘The battery is dead’ and ‘I have reason to believe that the battery is dead’. It is only in exceptional cases that the levels ever become mixed up: for example if the domain that one is reasoning about includes one’s own activities as a reasoner. In that case what is expressed by the statement ‘I have reason to believe that the battery is dead’ may be part of the raw data of the theory, something whose causes and effects are in question. This might be the case if one’s reasoning were concerned with cause and effect in the psychological domain. Ordinary causal reasoning, however, is not like this: Shoham’s motor-car example and my doorbell example are no exceptions.

Indeed, it is no doubt significant that Shoham does not allow his causal theory to contain ‘naked’ atomic sentences, i.e., sentences unadorned by any modal operator; and if any such were asserted, the theory would be unable to do anything with them. It follows that the expressions ϕ and $\Box\phi$ have very different statuses in the theory, so much so that it becomes rather doubtful whether they should both be regarded as expressions on the same footing at all. This applies whether one reads $\Box$ as ‘I have reason to believe that . . . ’, as I have suggested, or in one of the ways that Shoham hints at.

Moving on to item (b), we note that iterability of $\Box$ will only be possible if ϕ and $\Box\phi$ belong to the same syntactic category, for only in that case will $\Box\phi$ be admissible as an operand to $\Box$, allowing the formation of $\Box\Box\phi$. We have already seen reason to doubt whether this condition is fulfilled. For further evidence, though, we need only ask whether we ever need to consider propositions of the form

I have reason to believe that I have reason to believe that ϕ.

I think that intuition prompts us to answer ‘no’ here!

Actually, there are two distinct questions to consider. One is whether our syntax should allow such iteration; the other is whether such iteration, granted that it is allowed syntactically, is to have any semantic significance. By assuming that $\Box$ is a modal operator obeying the logic S5, Shoham in effect answers the former question positively and the latter negatively; for S5 is the modal logic in which all iterated strings of modal operators reduce to their final element, so that $\Box\Box\phi$ and $\Diamond\Box\phi$ are both equivalent to $\Box\phi$, and $\Box\Diamond\phi$ and $\Diamond\Diamond\phi$ to $\Diamond\phi$. Admittedly, Shoham concedes that the correct logic for his $\Box$ operator might be something a little weaker that S5, but what is significant for us is that this doesn’t actually make any difference to Shoham’s theory as he expounds it. The reason it doesn’t make any difference is that *Shoham never actually iterates any modal operators*. (Note in passing that Geffner explicitly disallows iteration in the case of his causal operator C: he rightly perceives that in ordinary causal reasoning we are hardly likely to want to say ‘we can explain why we can explain why . . . ’.)

We are thus left with (c) as the one respect in which ‘I have reason to believe that . . . ’ resembles a modal operator. For it does indeed fail to commute with negation. It is one thing not to have reason to believe that the battery is dead; it is quite another to have reason to believe that the battery is not dead; and this is the main justification we have for representing these two propositions in the forms $\Diamond\neg\phi$ (i.e., $\neg\Box\phi$) and $\Box\neg\phi$ respectively.

Is causality an epistemic notion?

A key idea in Shoham's theory is that *causality is an epistemic notion*. That is, what counts as the cause of an event depends on the causal laws by which the event is made predictable, and these causal laws in turn depend on the overall organization of the knowledge-base. Shoham says

> . . . A causes B if, whenever one believes A, and one does *not* believe that the background assumptions C are false, then one believes B.
>
> (Shoham 1990, p. 225)

As already remarked, this is much too psychological to suffice as a true account of causality; at the very least, we should replace 'believes' by 'has reason to believe'. But even then, the epistemic component would appear to be a superficial gloss. Surely what we really want is something like

> A causes B if, whenever A is the case, and the background assumptions C are true, then B is the case,

in which epistemic notions do not figure at all. This allows causal relations to be objective, whereas Shoham's theory reduces them to a subjective matter of propensities for belief.

To clarify this issue, let us draw as sharply as possible the contrast between two extreme views on the relationship between causality and human knowledge.

On the one hand, we have what may be labelled the *objectivist* view, according to which causality is a phenomenon that exists objectively, independently of what anyone says or believes. On this view, there could be, in principle, a 'true theory of the causal system of the world', encompassing all those causal statements which are, as a matter of fact, true.

It is quite consistent with the objectivist view that our ability to reason about the world is limited by the particular epistemic perspective we bring to it. Our knowledge is always incomplete, our generalizations never unexceptionable: hence our reasoning habits include various features, notably non-monotonicity, which help us to cope with our own limitations. These features apply to causal reasoning as much as to any other kind, but causal reasoning is not specially singled out as inherently non-mootonic.

Contrast with this the *subjectivist* view (essentially that advocated by Shoham), according to which the world in itself does not possess a causal structure, our idea of causality arising rather from our particular epistemic relation to reality. As such, causality will inherently partake of those features of human reasoning, such as non-monotonicity, which derive from that epistemic relationship. On this view, to try to expunge non-monotonicity from our account of causal reasoning would be like trying to expunge perspective from an account of human vision.

On either of these two views, causal reasoning will doubtless come out as non-monotonic, but on the objectivist view this is because *reasoning*, as actually practised, is always non-monotonic, whereas on the subjectivist view it is because cause is itself an epistemic notion and hence partakes of the non-monotonicity arising from our epistemic perspective.

Prima facie, the objectivist view is the natural one to go for in a computational model of causal reasoning. We do normally think of the causal relation as existing "out there" in the world. Shoham's advocacy of the subjectivist view rests mainly on the notion of a 'causal field', with the resulting non-monotonicity of causal reasoning.

Is causality inherently non-monotonic?

Shoham suggests that causal reasoning is a form of non-monotonic reasoning. This would imply that if we had perfect knowledge, and hence had no need for non-monotonicity in our reasoning, then we should have no need for causality either.

I do not want to deny that non-monotonicity is a feature of reasoning in general, and hence of causal reasoning in particular. For that matter, it is also a feature of reasoning about the weather, but this cannot in itself justify a claim that meteorological reasoning is a form of non-monotonic reasoning, for such a claim distinctly suggests that the weather itself is a non-monotonic phenomenon—whatever that could mean. Shoham's case for building non-monotonicity in as an intrinsic component of causal reasoning rests on the following considerations.

Shoham asks us to consider his motor-car example. Normally, we say that turning the key causes the motor to start, even though what actually enables the motor to start is a conjunction of circumstances including not just turning the key but also the battery's being charged, and connected up, etc. We don't say that *these* are what cause the motor to start, though—we would not normally say, for example, in answer to the question 'Why did the motor start?', 'Because the battery was connected up'. These are things which we tend to assume by default: they are the normal background conditions (the "causal field"), which enable the motor to be started, but do not in themselves suffice to start it.

Shoham now asks us to envisage that the key jams in the ON position, so that in order to stop the motor we have to disconnect the battery. We can't afford to get the key fixed, so from now on we start the car by connecting up the battery. Shoham says: 'After a short while it will seem natural to say that it is the connecting of the battery that causes the car to start'. And thus he rests his case: change the background assumptions, and you change the pattern of causal reasoning, in particular you change what *counts* as a cause. In practical terms, this means that an addition to the background conditions can result in a previously acceptable causal inference becoming no longer acceptable: in other words that causal inference is non-monotonic.

I believe that there has been a sleight of hand here. Shoham has tried to present us with a puzzle: why is it, he seems to ask, that normally we would say that the cause of the motor's starting is the ignition key coming to the ON position, and not the battery's being connected, whereas in the "jammed key" scenario, we *do* say that the connection of the battery is the cause of the motor's starting, rather than the key's being in the ON position? He uses this example to motivate the idea of a causal field, without which the thing would be—so he would have us believe—inexplicable.

But surely he has missed something very obvious? The thing that we call the cause is in both cases *something that happens*, i.e., an event; whereas the causal field is composed of *states of affairs*, i.e., the states which obtain at the time that the event happens. The sleight of hand comes from exploiting the ambiguity of "the connection of the battery" as between a state-reading and an event-reading, i.e., between the circumstance of the battery's being in a connected state (which cannot be a cause, only a condition for a cause to be effective), and the event of the battery's coming to be connected (which *can* be a cause).

All this may not amount to anything more than a demonstration that Shoham's example is ill-chosen. Perhaps a better example could be found in which the cause and the causal field conditions are all events, or all states. This needs further investigation. But it does show that one cannot be too careful about respecting distinctions of *aspectual character*, such as that between states and events (Galton 1984)—which Shoham explicitly disavows when he says 'I will not introduce a distinction among events, facts, processes, and so on, at the primitive level'.

Of course, even if my criticism of Shoham is right, this does not mean that non-monotonicity is necessarily absent from causal reasoning. What it means, though, is that Shoham is wrong to locate the source of this non-monotonicity in an intrinsic property of causal reasoning as such, as opposed to reasoning quite generally. It may be that non-monotonicity, as a feature of everyday reasoning, is all-pervasive, and hence interacts with causal reasoning as with any other form of reasoning; where I believe Shoham to be mistaken is in supposing that there is something about the notion of cause which is inextricably bound up with non-monotonicity.

Concluding remarks

Shoham's causal theory rests on two basic assumptions which we have called into question in this paper. They are

1. Causality is an epistemic notion;

2. Causal reasoning is non-monotonic.

These assumptions are elaborated by means of a number of subsidary assumptions, notably that both epistemic notions and non-monotonicity can or should be handled by means of modal logic.

In this paper I have argued that causality is not intrinsically epistemic, nor causal reasoning non-mononotonic. In so far as epistemic and non-monotonic notions are implicated in our treatment of causality, it is because they are features of reasoning generally. My claim is, then, that causality has no special affinity with either epistemic notions or non-monotonicity. We might want both of these features anyway in order to handle default reasoning, incremental knowledge growth, and so on, but whatever mechanism we employ for it does not have to be particularly bound up with the causal reasoning mechanism.

Indeed, I believe it makes for much greater conceptual clarity if we strive as far as possible to keep the mechanics of reasoning separate from the subject-matter that we are reasoning about[2]. If causality is truly objective—and it had better be, if anything at all is—then a causal law ought to have some such form as

$$\phi_1 \wedge \phi_2 \wedge \cdots \wedge \phi_m \wedge \chi_1 \wedge \chi_2 \wedge \cdots \wedge \chi_n \rightarrow \psi.$$

Here the ϕ_i are events, the χ_j are states; if the states form a causal field against which the events have their effects, this is because the methods of reasoning we apply will make default assumptions about the states but not about the events. Thus an adequate theory of causal reasoning must, in my view, be able to make a clear distinction between states and events; it will also need to give an account of how these categories are related to processes. In short, causal reasoning must be sensitive to aspectual character.

The view I have expressed here is more in harmony with the method advocated by Lifschitz (1987), in which non-monotonicity arises not from the way in which the causal relation is described (not, in other words, from the declarative logic of the theory) but from the general principles of reasoning, specifically circumscription, which are brought to bear on it. Lifschitz's theory makes use of the Situation Calculus (McCarthy and Hayes 1969), which does incorporate some aspectual distinctions lacking from Shoham's account, as embodied in the categories of *action* and *fluent*: only an action can be a cause, and only a fluent can be a precondition for action. My main objection to the Situation Calculus and, by implication, to Lifshitz's use of it, is that it reifies *types* of state and event, whereas I believe that it is more satisfactory to reify *tokens*. For a discussion of this issue, see

[2]Just as, traditionally, logic has always been regarded as *transcendent*, that is, independent of this or that specific subject-matter.

(Galton 1991).

Acknowledgements. I should like to thank John Gooday, Yoav Shoham, and the two anonymous referees for their careful and constructive comments on the original version of this paper.

References

Galton, A.P. 1984. *The Logic of Aspect*. Oxford, UK: Clarendon Press.

Galton, A.P. 1991. Reified temporal theories and how to unreify them. To appear in Proceedings of the Twelfth International Joint Conference on Artificial Intelligence.

Geffner, H. 1990. Causal theories for non-monotonic reasoning. In Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90), 524-530. Menlo Park, Calif.: AAAI Press/MIT Press.

Lifschitz, V. 1987. Formal theories of action (Preliminary report). In Proceedings of the Tenth International Joint Conference on Artificial Intelligence, 966-972. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, Inc.

McCarthy, J. and Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Melzer, B. and Michie, D. (eds.), *Machine Intelligence 4*: 465-502. Edinburgh: Edinburgh University Press.

Pearl, J. 1988. Embracing causality in default reasoning. *Artificial Intelligence* 35(2): 259-271.

Shoham, Y. 1988. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence*. Cambridge, Mass.: The MIT Press.

Shoham, Y. 1990. Nonmonotonic reasoning and causation. *Cognitive Science* 14: 213-252.