Counterfactuals for Causal Responsibility in Legal Contexts

Holger Andreas[1], Matthias Armgardt[2], and Mario Günther[3]

[1]University of British Columbia
[2]University Hamburg
[3]Australian National University

Penultimate draft of a paper to appear in *Artificial Intelligence and Law*

Abstract

We define a formal semantics of conditionals based on *normatively ideal worlds*. Such worlds are described informally by Armgardt (2018) to address well-known problems of the counterfactual approach to causation. Drawing on Armgardt's proposal, we use iterated conditionals in order to analyse causal relations in scenarios of multi-agent interaction. This results in a refined counterfactual approach to causal responsibility in legal contexts, which solves overdetermination problems in an intuitively accessible manner.

1 Introduction

We define a formal semantics of conditionals based on *normatively ideal worlds*. Such worlds are described informally by Armgardt (2018) to address well-known problems of the counterfactual approach to causation. In essence, a possible world is normatively ideal iff all agents act according to their legal duties. Normatively ideal worlds are a promising tool to model the violation of legal duties and the ascription of causal responsibility that goes along with it. We refine Armgardt's proposal in order to better account for individual causal responsibility in scenarios of multi-agent interaction.

The main objective of our investigation is to solve problems of overdetermination in an intuitively accessible manner within the counterfactual approach to causation. Scholars of law have become well aware of such and related problems (see, e.g., Moore (2009, Ch. 17)). A standard example goes as follows. Suppose two arsonists set a house on fire simultaneously. The fire destroys the house completely. Further, suppose the actions of each arsonist is sufficient to burn down the house. Then, intuitively, we consider both arsonists causally responsible for the damage to the house. However, the simple counterfactual test for causation fails to capture this intuition. This test says that an occurring event causes another just in case, had it not occurred, the other would not have occurred as well. Overdetermination is a counterexample to the simple counterfactual test. Had one of the arsonists not set the house on fire, the house would still have burnt down – due to the action of the other arsonist.

Why do we aim at a simple and intuitively accessible solution to the overdetermination problem in the context of law? This objective is motivated by two observations. First, counterfactual reasoning is well established among both scholars and practitioners of law. In the context of law, the simple counterfactual test for causation is often referred to as the *but-for test*: but for a certain action, a specific damage or harm would not have occurred. If an action passes this test, it is deemed causally responsible for the damage or harm. The basic idea of *but-for causation* is also captured by the notion of *conditio sine qua non*: an action is considered causally responsible for a damage or harm if this action is a condition without which the damage or harm would not have been done. Our proposal will preserve the tradition of counterfactual reasoning in law.

Second, extant solutions to the overdetermination problem within the counterfactual approach are cumbersome and difficult to translate without using a number of technical terms from set theory and causal models. As noted by Liepiņa et al. (2020), the updated Halpern-Pearl definition of causation in Halpern and Pearl (2005) and Halpern (2016) exceeds the comfort level of most scholars and practitioners of law. The modified Halpern-Pearl definition in Halpern (2015) and Halpern (2016) is simpler and more accessible, but fails to capture individual causes in scenarios of symmetric overdetermination.[1] Our solution to the overdetermination problem is both conceptually and syntactically simpler than the modified Halpern-Pearl definition in Halpern (2015) and Halpern (2016).

[1] On the latter definition, the set of overdetermining causes qualifies as a cause, but the members of such a set do not (Andreas and Günther 2021).

Woods (2018) observed two potential problems with Armgardt's account of causation using normatively ideal worlds. First, by omitting any explicit similarity ordering among possible worlds, the actual world drops out of the picture in Armgardt's informal semantics of counterfactuals. This semantics may therefore fail to evaluate counterfactual conditionals relative to the actual world. In the eyes of Woods (2018, p. 713), "it is not yet clear how the NIW model can preserve the counterfactuality of the but-for rule without relying on maximal similarity assumptions". (NIW stands for the concept of normatively ideal worlds.) Second, Armgardt's account is not sufficiently fine-grained to recognize individual causal contributions in scenarios where the damage is overdetermined, while individual contributions differ among one another. Both of these two problems are addressed in the analysis of causal responsibility suggested below.

2 Semantics

Suppose we have a causal scenario in which certain events occur and certain actions are performed. In the context of law, we have to deal with causal scenarios where certain legal duties are violated.[2] And we have to investigate and to determine the causal consequences of such violations. Armgardt (2018) invokes *normatively ideal worlds* in order to carry out this determination. In brief, the idea is that a violation of a law affects an actual event e if and only if e does not occur in the normatively ideal worlds of the corresponding causal scenario. A world is normatively ideal if and only if all agents involved in the causal scenario act according to their legal duties. The following type of counterfactual is suggested for determining causal responsibility in legal contexts (Armgardt 2018, Sect. 5.1):

> If all involved agents had acted according to their legal duties, would then the harm have not occurred?

How can we supplement these ideas with a formal semantics? Let P be a set of propositional variables such that P captures all the events of the causal scenario in question. That is, we have for every event – occurrent or not – a corresponding variable $p \in P$. p means that the event occurs, while $\neg p$ says that the event does

[2] This observation has already been made by Hanau (1971), and more recently by Schaffer (2010) among others. Armgardt (2018) goes beyond Schaffer by considering the behaviour of a group of agents in terms of possible worlds.

not occur. Note that we use the term 'events' in a broad sense so as to include non-occurrences of events, or absences. Further, let $\mathcal{L}(P)$ be a propositional language of classical logic, based on P. $\mathcal{L}(P)$ contains all the propositional formulas built using the standard propositional connectives $(\neg, \wedge, \vee, \rightarrow, \leftrightarrow)$ and the members of P.

Obviously, possible worlds are needed in a framework of normatively ideal worlds. Let W be the set of possible worlds that may be described using the language $\mathcal{L}(P)$. $V : P \rightarrow \mathcal{P}(W)$ is a valuation function that tells us which propositions are true in which possible worlds. To be more precise, $V(p)$ gives us the set of possible worlds in which the proposition $p \in P$ is true. Mathematically, V is a mapping of the set P onto the powerset of the set W of possible worlds.

Suppose we know what it means for an agent to act – in a given causal scenario – according to his or her legal duties. Then, there is a well-defined set N of normatively ideal worlds such that $N \subseteq W$. However, not all of these worlds need to be relevant for determining the causal consequences of unlawful behaviour. In line with Lewis's (1973b) semantics of counterfactuals, we consider those possible worlds w most relevant that are a member of N and that minimally deviate from our actual world w_0. For this to be made precise, we need to introduce a few more formal concepts.

We will "measure" the degree of similarity between two worlds in terms of the sets of *literals* that the two worlds respectively satisfy. (A propositional literal is an atomic propositional formula or the negation of such a formula.) Let $L(w)$ be the set of *literals* that are true in the possible world w. $L(w)$ is defined as follows:

$$L(w) = \{p \mid p \in P \text{ and } w \in V(p)\} \cup \{\neg p \mid p \in P \text{ and } w \notin V(p)\}. \qquad \text{(L(w))}$$

Finally, we assume that in a normatively ideal world all presumed laws of nature are satisfied. This assumption allows for a simpler definition of relevance among possible worlds to be introduced shortly. Of course, we must wonder whether there are cases where laws of nature make it impossible to act according to one's legal duties. We discuss this problem in the next section, and indicate how the assumption in question may be lifted.

We are now in a position to introduce an ordering $<$ of relevance among possible worlds:

Definition 1. $w < w'$
Let w_0 be the actual world of the causal scenario considered. We say that the

possible world w is more relevant than w' and write $w < w'$ if and only if $w \in N$ and

(1) $w' \notin N$, or

(2) $|L(w) \cap L(w_0)| > |L(w') \cap L(w_0)|$.

That is, $w < w'$ obtains if and only if (i) w is a normatively ideal world, while w' is not, or (ii) w is a normatively ideal world and more similar than w' to the actual world w_0. Condition (2) simply says that the set of literals satisfied by both w and w_0 is "larger" than the corresponding set for w' and w_0 is. In other words, the greater the intersection of the sets $L(w)$ and $L(w_0)$ is, the more similar are the worlds w and w_0. A set is larger than another if and only if it has a greater cardinality; $|A|$ designates the cardinality of set A.

The minimum of the ordering thus defined gives us the possible worlds that are most relevant for counterfactual considerations where all agents act according to their legal duties.

$$<_{min} = \{w \mid \text{ there is no } w' \text{ such that } w' < w\} \qquad (<_{min})$$

The idea underlying this definition is straightforward: a world w is most relevant if and only if there is none that is more relevant than w.

We are now in a position to define a counterfactual conditional with the following meaning: Had all agents acted according to their legal duties, ϕ would have been true:

$$LD \boxdot\!\!\rightarrow \phi \text{ iff } w \models \phi \text{ for all } w \in <_{min}$$

where *LD* stands for the antecedent that all agents act according to their legal duties. This definition says that $LD \boxdot\!\!\rightarrow \phi$ is true if and only if ϕ is true in all possible worlds that are most relevant in the sense of the relevance ordering $<$.

Applied to singular events e, we can say that a group is legally, and causally, responsible for the occurrence of e if and only if (i) $w_0 \models e$ and (ii) $LD \boxdot\!\!\rightarrow \neg e$. In words, e must be an actual event that would not have occurred if all agents of the group had acted according to their legal duties. This is the core of Armgardt's (2018) analysis of causal responsibility in legal contexts. We have now complemented this analysis by a similarity ordering of possible worlds, called *ordering*

of relevance. Woods's worry – that it might be premature to abandon the similarity semantics of counterfactuals altogether – is thereby addressed (Woods 2018, p. 713).

Why is an ordering of relevance, or similarity, needed? Suppose an arsonist burns down a house. In a normatively ideal world, this person would not have started a fire to burn down the house. However, strictly speaking, it is not correct to say that the house does not burn down in all possible worlds in which all agents involved act according to their legal duties. For among these worlds there are also ones in which the house burns down because of events other than the action of an arsonist. For example, there are normatively ideal worlds in which the house burns down because of lightning, a short circuit, or a sudden heart attack of the home owner while cooking. Only in those normatively ideal worlds that are closest to our actual world it holds true that the house does not burn down. The relevance ordering $<$ makes sure we capture those possible worlds when evaluating conditionals.

3 Open Problems

In the previous section, we have furnished Armgardt's proposal with a formal similarity semantics of counterfactual conditionals. Thus we have addressed one of the two problems which Woods (2018) observed for this proposal. The other problem remains to be solved: within a group of agents, the individual causal contributions may differ, and eventually we want to ascribe causal responsibility to individuals. Let us take a closer look at this problem.

The core account in Armgardt (2018) ascribes causal responsibility to a group of people. As explained in the previous section, it is guided by the following counterfactual consideration (Section 5.1):

> If all involved agents had acted according to their legal duties, would then the harm have not occurred?

Working with counterfactuals that concern a whole group of agents ("all involved agents") is primarily motivated by problems of overdetermination and preemption. In such scenarios, our counterfactual considerations need to go beyond the actions of just a single agent. Take the example of a house set on fire, independently and simultaneously, by two people. By assumption, the action of each agent suffices to burn down the house. So the house would have burnt down, even if one of the

agents had not set the house on fire. And yet, we consider the individual action of setting the house on fire a cause of the house burning down. Armgardt's analysis allows us to recognize this causal judgement since the house would not have burnt down if the two agents had acted according to their legal duties.

However, it does not seem to be always true, as Armgardt himself admits, that "all agents involved" are causally responsible for an event that is caused by a group of agents. Take, for example, the poor homeowner who is sitting in the backyard when his or her house is set on fire. He or she is certainly "involved", but not in the manner that his or her actions contribute to the home burning down.

Furthermore, we may have different types of unlawful behaviour involved in one and the same causal scenario. If so, we may want to differentiate between these types, and assess their contribution to the harm or damage in question. To expand our example, suppose the firefighters did not show up within a reasonable period of time for reasons they are responsible for. If they had, they would have been able to extinguish the fire before the house burnt down completely. So, the fire damage could have been prevented, in part, by diligent behaviour of the firefighters. Clearly, setting a house on fire and coming too late to extinguish a fire are two different types of misconduct that should be distinguished from one another. It is necessary to judge these two types of unlawful behaviour separately.

More generally, we can say that causal responsibility must – in legal contexts – be eventually ascribed to individuals. How can we specify the individual causal contributions within a whole group of agents? In the next two sections, we shall outline a proposal.

Finally, our formal account of Armgardt's proposal may be criticized for two reasons. First, strictly speaking, the sentence *LD* (saying that all agents act according to their legal duties) is not a sentence of the language $\mathcal{L}(P)$. At least, we have not said how this sentence could be spelled out in this language. However, this should not be difficult to achieve. A simple, though not very elegant way to express *LD* in $\mathcal{L}(P)$ would be based on the literals true in the normatively ideal worlds, respectively. $\bigvee_{w \in N} \bigwedge L(w)$ is a sentence of $\mathcal{L}(P)$ saying that we are in one of the normatively ideal worlds. ($\bigvee_{w \in N} \bigwedge L(w)$ designates some disjunction such that each disjunct is a conjunction of all the literals true in w ($w \in N$)).

A more elegant way of spelling out *LD* is to specify the legal norms – relevant to the causal scenario – in the language $\mathcal{L}(P)$. Extending the propositional language $\mathcal{L}(P)$ to a first-order language may be helpful for this. In the extended language, we could for example say that people do not set other people's houses on fire. *LD*

would then be a conjunction of norms, each of which is expressed by a statement that describes lawful behaviour.

Second, in the above formal account, we have assumed that the normatively ideal worlds are such that presumed laws of nature are satisfied. This is a reasonable assumption since in the context of law, presumed laws of nature are assumed to hold true, and we do not want to further investigate such laws. Explicit consideration of laws of nature may nonetheless be desirable for mainly two reasons. First, to deal with scenarios where it is, physically or practically, impossible to act according to one's legal duties. Second, in Lewis's (1973b) semantics, any counterfactual consideration involves "a small miracle" because we need to hypothetically change the course of events in our world. The last point, however, is not relevant if we work with "small worlds" that concern a given causal scenario and that lack an extended causal history. In such worlds, we ignore the causal history of the causes under consideration. Notably, this strategy is also pursued in the causal models semantics by Halpern (2000) and Halpern and Pearl (2005). In any case, it does not seem to be difficult to consider laws of nature within the above formal account. For example, we could require that, for a world w to be at least as relevant as world w', w must satisfy the laws of nature (relevant in the causal scenario) to at least the same degree to which w' does.

4 Iterated Difference-Making

Lewis (1973b) does not only study counterfactuals in our world, but also counterfactuals in possible worlds other than ours. Thereby, he succeeds in defining iterated conditionals and counterfactuals. To give a simple example of an iterated counterfactual conditional: if I had lived 100 years ago, then, had I been into climbing, I would have tried to do a first ascent at some mountain in the Himalaya. This counterfactual has the following logical form:

$$\alpha \mathrel{\Box\!\!\rightarrow} (\beta \mathrel{\Box\!\!\rightarrow} \gamma).$$

In words, if α were true, γ would be true if β were true. Such a conditional may or may not be equivalent with

$$\alpha \wedge \beta \mathrel{\Box\!\!\rightarrow} \gamma.$$

In their ongoing work, Mario Günther and Holger Andreas exploit iterated counterfactual considerations to address the infamous problem of overdetermination

and related issues. This account seems worth considering in light of the problems described in the previous section. Andreas and Günther (2021) begin with considering counterfactuals where a set of actual events is absent. In formal terms:

Definition 2. *c* **causes** *e* (preliminary)
c causes *e* if and only if there is a set C of actual events such that $c \in C$ and

(C1) *c* and *e* occur in the actual world, and

(C2) $\neg \bigvee C \mathrel{\Box\!\!\rightarrow} \neg e$.

This condition simply says that *e* would not occur if none of the events of C were to occur. For simplicity, we use the symbols *c* and *e* to designate both events and sentences that assert the occurrence of corresponding events. Likewise, C is used for a set of events and a set of sentences about the occurrence of events. $\bigvee C$ designates some disjunction of the members of C. (If $C = \{c\}$, $\bigvee C$ is given by *c*.)

Definition 2 is a preliminary analysis of causation that provides us with a solution to the problem of overdetermination. Applied to our example: had none of the arsonists set the house on fire, the house would not have burnt down. Hence, the single action of one of the arsonists is a member of a set C such that condition (C2) is satisfied. Since condition (C1) is satisfied as well, this action is a cause of the house burning down.

However, this solution is too simple. Causally non-relevant events qualify as causes on the analysis by Definition 2. Take the following set C of events:

$$C = \{\text{arsonist}_1 \text{ sets house on fire},\ \text{arsonist}_2 \text{ sets house on fire},\ \text{homeowner sits in the backyard}\}.$$

Then the event of the homeowner sitting in the backyard comes out as a cause of the house burning down. For this event satisfies the conditions (C1) and (C2) for the set C.

At this point, iterated counterfactuals come into play. Andreas and Günther (2021) require that – in the counterfactual worlds where none of the events of C occurs – each of the non-occurrences of C must make a difference as regards the non-occurrence of *e*. In formal terms:

$$\text{for any non-empty set } C' \subset C, \neg\bigvee C \mathrel{\Box\!\!\rightarrow} (\bigwedge C' \mathrel{\Box\!\!\rightarrow} e). \tag{C3}$$

This condition makes use of an iterated counterfactual. It says that, in the counterfactual worlds where none of the events in C occurs, if some events of C were

to occur after all, then e would occur. Taken together, (C2) and (C3) say that the putative effect e is absent *only* on condition that none of the events in C occurs. Otherwise e occurs. This is the core account:

Definition 3. c **causes** e
c causes e if and only if there is a set C of actual events such that $c \in C$ and

(C1) c and e occur in the actual world,

(C2) $\neg \bigvee C \mathrel{\Box\!\!\rightarrow} \neg e$, and

(C3) for any non-empty set $C' \subset C$, $\neg \bigvee C \mathrel{\Box\!\!\rightarrow} (\bigwedge C' \mathrel{\Box\!\!\rightarrow} e)$.

Let us exemplify the basic idea of this account with our running example. c_1 and c_2 may designate the two independent actions of setting the house on fire. Then, we have

$$\neg c_1 \mathrel{\Box\!\!\rightarrow} e \tag{1}$$

$$\neg c_2 \mathrel{\Box\!\!\rightarrow} e. \tag{2}$$

That is, had one of the arsonists not set the house on fire, the house would still have burnt down. So the simple counterfactual account does not work. However, the iterated counterfactual account does. For the following counterfactuals come out true (in the actual world):

$$\neg(c_1 \vee c_2) \mathrel{\Box\!\!\rightarrow} \neg e \tag{3}$$

$$\neg(c_1 \vee c_2) \mathrel{\Box\!\!\rightarrow} (c_1 \mathrel{\Box\!\!\rightarrow} e) \tag{4}$$

$$\neg(c_1 \vee c_2) \mathrel{\Box\!\!\rightarrow} (c_2 \mathrel{\Box\!\!\rightarrow} e). \tag{5}$$

By (3), we know that condition (C2) of the refined account is satisfied. To be more precise, (C2) is satisfied for $C = \{c_1, c_2\}$. By (4) and (5), we know that condition (C3) of the refined account holds for $C = \{c_1, c_2\}$. $C' = \{c_1\}$ is a non-empty, proper subset of $C = \{c_1, c_2\}$ such that (C3) is true. Likewise, for $C' = \{c_2\}$. Condition (C1) is obviously true as well.

Now, let s designate the event that the homeowner is sitting in the backyard. The following iterated counterfactual is then true:

$$\neg(c_1 \vee c_2 \vee s) \mathrel{\Box\!\!\rightarrow} (s \mathrel{\Box\!\!\rightarrow} \neg e). \tag{6}$$

And so the following iterated counterfactual *fails* to hold true:

$$\neg(c_1 \vee c_2 \vee s) \; \Box\!\!\rightarrow (s \; \Box\!\!\rightarrow e). \tag{7}$$

Hence, the set $C = \{c_1, c_2, s\}$ fails to satisfy condition (C3). The iterated counterfactual analysis does not recognize the event of the homeowner sitting in the backyard as a cause of the house burning down, as it should be.

Condition (C3) may be described as a condition of relevance. It excludes that events that are causally non-relevant for the effect e come out as causes since they satisfy conditions (C1) and (C2). Without (C3) any causally non-relevant event would qualify as a cause.

Notably, the condition of relevance could also be expressed as follows:

$$\text{for any } C' \subset C, \neg\bigvee C' \; \Box\!\!\rightarrow e. \tag{C3$'$}$$

That is, if only the events of a proper subset of C are absent, the effect e still occurs. Applied to our example, if only one of the arsonists does not set the house on fire, the house still burns down since the other arsonist remains active. $\neg \bigvee C'$ stands for the negation of some disjunction of the members of C'. $\neg \bigvee C'$ says that all members of C' are false. For this to be seen more clearly, recall that $\neg(C_1 \vee \ldots \vee C_n)$ is logically equivalent to $\neg C_1 \wedge \ldots \wedge \neg C_n$ (according to the DeMorgan laws of propositional logic). There thus is a way to express the relevance condition without iterated counterfactuals. But the motivation for (C3$'$) is aptly explained by considerations about iterated difference-making. This is why we started with the formulation in terms of iterated counterfactuals.

It is worth noting that the refined counterfactual account is conservative with respect to the simple counterfactual account of causation. Suppose c qualifies as a cause of e according to the simple account. We can prove that c continues to be a cause of e on the refined account (which is given by Definition 3). Simply take $C = \{c\}$. (C2) is then equivalent to $\neg c \; \Box\!\!\rightarrow \neg e$. And (C3) is trivially satisfied since there is no non-empty, proper subset of $\{c\}$. Hence, the iterated account captures all genuine causal relations captured by the simple counterfactual account of causation.

Iterated difference-making identifies genuine causes in scenarios of overdetermination. c is a cause of e if and only if there is a set C of actual events such that two conditions are satisfied. First, the whole set C of events makes a difference as to whether e occurs – in the sense that e would be absent if none of the events of C were to occur. Second, in the counterfactual worlds where none of the events

of C occurs, each of the non-occurrences of C is a difference-maker as regards the non-occurrence of e. Were some event of C to occur in such a counterfactual world after all, the putative effect would occur as well, provided $c \in C$ is a cause of e.

5 Normatively Ideal Worlds Continued

It seems as if ideas about iterated difference-making allow us to solve overdetermination problems in an intuitively accessible manner. So one might wonder why we still need normatively ideal worlds in an account of causal responsibility. Well, consider this case: Bob promises to water Ann's plant, doesn't water it, the plant dries up and dies. If Bob had watered the plant, the plant would not have died. Common sense has it that Bob's *not* watering the plant caused its death.

Beebee (2004) and Moore (2009) disagree: Bob's omission to water the plant is no cause. They argue that absences cannot be causes because causation is a relation between events and absences are no events. Lewis (2000, 2004), by contrast, argues that absences can be causes because they can seamlessly enter into patterns of counterfactual dependence. We accept Lewis's argument for the sake of preserving the difference-making idea behind counterfactual accounts. Some absences and omissions make a difference, and so should come out as causes on a "broad and nondiscriminatory concept of causation" (Lewis 1973a, p. 559). Hence, we use the term 'event' in a broad sense including absences. We should add that we likewise use the term 'action' in the wider sense which includes omissions.

Our analysis says that Bob's failure to water the plant caused its death – like all accounts that treat events and absences on a par and that take counterfactual dependence between actual events and absences to be sufficient for causation. So far so good.

However, if Bob's omission to water the plant counts as a cause of the plant's death, then there is a lot of causation by omission. Carla omitted to water the plant, and so did Queen Elizabeth II. If either of them would have watered the plant, the plant would not have died. True. And yet, we would refrain from saying that Carla or the Queen are causally responsible for the plant's death. After all, they did neither promise to water the plant nor had they any other obligation to do so.

The basic idea is that you are only causally responsible for an outcome if you violate some norms. Unlike Bob, neither Carla nor the Queen break a promise to water the plant. So it is Bob – not Carla and not the Queen – who violates a promise

by not watering the plant. In any normatively ideal world, Bob keeps his promise by (not not) watering the plant. By contrast, there are normatively ideal worlds, in which Carla and the Queen act according to their duties and do not water the plant. As they made no promise to water the plant, they have no corresponding duty. This is how we can model the violation of norms by normatively ideal worlds.

Here is a first stab at merging ideas about iterated difference-making with Armgardt's (2018) account. Let $N(G)$ be the set of normatively ideal worlds where all members of the group G act according to their legal duties, and $LD(G)$ be the proposition saying that all members of G act according to their legal duties. We can then state the following account of causal responsibility for individuals in legal contexts:

Definition 4. Causal Responsibility of Individuals
An individual i is causally responsible for an actual event e if and only if there is a group G of people such that $i \in G$ and

(R1) $LD(G) \boxdot\!\!\rightarrow \neg e$, and

(R2) for any $G' \subset G$, $LD(G') \boxdot\!\!\rightarrow e$.

The second condition ensures that innocent people are not considered legally responsible for a damage. For simplicity, we have modelled (R2) in the image of (C3′) rather than (C3). Applied to our running example: condition (R1) is true for G being both {arsonist$_1$, arsonist$_2$, homeowner} and {arsonist$_1$, arsonist$_2$}. But condition (R2) fails to hold for {arsonist$_1$, arsonist$_2$, homeowner }, while this condition is satisfied for {arsonist$_1$, arsonist$_2$}. In brief, the group {arsonist$_1$, arsonist$_2$} passes the conditions of the definition, while the larger group does not.

Recall that the conditional $\boxdot\!\!\rightarrow$ is defined by a relevance ordering, the definition of which was given in Section 2. This definition remains in place, with the qualification that the sets $N(G)$ and $N(G')$ of normatively ideal worlds are relativized to the actions – including omissions – of the members of a group of agents. They may be further relativized to sets of actions of whatever individuals. This is necessary when a single agent violates different legal duties such that these violations have different causal consequences.

Notably, in the absence of overdetermination and related problems, G is simply the singleton of the individual i, i.e, $G = \{i\}$. In this case, it suffices to consider the possible worlds where the defendant acted according to his or her legal duties. Technically speaking, the present proposal contains the simple but-for test

for causal responsibility as a special case. In cases where we have causation and the simple but-for test works in the intended way, our refined account of causal responsibility agrees with this test. This is good news if one thinks that this test is worth preserving and refining.

Moreover, we can merge the definition of causation in terms of iterated difference-making directly with considerations about normatively ideal worlds. This results in an account of causal responsibility for individual actions:

Definition 5. Causal Responsibility of Actions
An action a by an individual i is causally responsible for an actual event e if and only if there is a set C of actual events such that $a \in C$ and

(R1) $\neg \bigvee C \mathrel{\Box\!\!\rightarrow} \neg e$,

(R2) for any non-empty set $C' \subset C$, $\neg \bigvee C' \mathrel{\Box\!\!\rightarrow} e$, and

(R3) the individual i does not perform action a in any of the normatively ideal worlds.

Recall Bob who promised to water the plant, but failed to do so. Bound by his promise, Bob does (not not) water the plant in any normatively ideal world. His failure to water the plant is thus causally responsible for the plant's death. Carla and the Queen, on the other hand, did neither promise to water the plant nor had any other obligation to do so. That is, there are normatively ideal worlds, in which Carla and the Queen act according to their duties and do not water the plant. Hence, their omissions do not satisfy (R3) of Definition 5, and so are not causally responsible for the plant's death.[3]

Take an example closer to a legal context. Suppose Peter sees a child falling into a river. He does nothing to save the child, and so the child drowns. If Peter had jumped into the river to save the child from drowning, he would have done so without much effort. Peter's failure to provide assistance is thus a cause of the child's drowning. In any normatively ideal world, Peter does not fail to help the child. Hence, Peter's failure to give assistance is causally responsible for the child's

[3]We say that Bob's omission and the Queen's omission, respectively, are causes of the plant's death. But only Bob's omission is causally responsible for its death. McGrath (2005) has a stricter view on causation: an omission is a cause only if some norm is violated. She thinks Bob's failure to water the plant is a cause of its death, while the Queen's omission to water the plant is not. Note that she thereby does not preserve the verdicts of the simple counterfactual test even though she treats events and absences on a par.

death. Paul is far away and does not see the child falling into the river. He likewise does not jump into the river to save the child from drowning. Paul's omission is also a cause of the child's drowning on the simple counterfactual account. However, Paul is not even aware of the child. So there are normatively ideal worlds where Paul acts according to all his legal duties and yet the child drowns. Paul's omission is therefore not causally responsible for the child's death. Our analysis says all of this.

Let's return to overdetermination. Both definitions 4 and 5 are fine-grained enough to distinguish between individual causal contributions in scenarios where the damage is overdetermined, while individual contributions differ among one another. Moore (2009, p. 417n) refers to such scenarios as cases of *asymmetric overdetermination*. Suppose there are two arsonists who set a house on fire at the same time, but act independently. One wants to burn down the house completely, while the other only wants to cause some damage, without aiming at a complete destruction. Let's say one uses fire accelerator, while the other does not. Because of the use of fire accelerator, the house burns down completely. Otherwise the damage would have been partial and repairable. The fire would have damaged the interiors and some drywall, but not the frame and the roof of the house.

In this scenario we have to consider two different effects: first, complete destruction of the house, and second, damage to the interiors and drywall. The first damage is not overdetermined. So it does not pose specific difficulties. The damage to interiors and drywall is overdetermined. If the house burns down completely, then certainly interiors and drywall are damaged as well. Let us designate the arsonist with fire accelerator i_1, while the other is designated by i_2. Further, let e be the event that interiors and drywall are damaged. Then, $G = \{i_1, i_2\}$ satisfies Definition 4 relative to the event e. So, the arsonist without the fire accelerator is causally, and legally, responsible for the damage to interiors and drywall. But he or she is neither causally nor legally responsible for the complete destruction of the house. In a similar vein, Definition 5 allows us to distinguish between the individual causal contributions of the two arsonists.[4]

[4]Braham and van Hees (2009) have shown how degrees of causal contribution may be formally distinguished using what is called the *NESS test* for causation. NESS stands for *necessary element of a sufficient set*, sufficient to bring about a certain effect. Our motivation for a counterfactual approach to distinguishing individual causal contributions is rooted in the observation that the but-for test for causation is better known among scholars and practitioners of law than the NESS test.

6 Collective Actions

Let us finally apply our analysis of causal responsibility to collective actions. We show that this analysis is fine-grained enough to capture causal relations of, at least, some collective actions. Suppose the board of directors of a company receives reports which indicate that one of the company's products is toxic. The board has to decide whether sales of the product should be stopped until further studies disconfirm that the product is toxic. Each board member is legally obliged to do so. However, the majority of the board members are more concerned with their annual bonus payments (which depend on the company's revenue) than with the long-term prosperity of the company and the health of their customers. And they assume that empirical evidence of toxicity will anyways not be leaked to the public. Hence, the majority of the board members decides not to stop manufacturing and selling the product in question. To make our example more concrete, assume the board has eleven members and nine members were in favour of continuing with sales and production. For putting sales and production on hold, simple majority of the votes was needed.

After years of successful sales, the product turns out to be actually toxic. A significant number of customers reported health problems, which could be traced back to the product in question by well-validated scientific studies. Obviously, we should consider each board member who decided to ignore the early reports causally, and legally, responsible for the health issues. However, the simple but-for test fails to deliver this result because roughly 82 percent of the board members were in favour of ignoring the early reports and simple majority of the votes was needed to stop production and sales. Using the simple counterfactual test, each member who decided to ignore the early reports could argue that his or her vote was not decisive for the company's decision to not stop sales and production. This type of excuse is, of course, not acceptable.

Our refined counterfactual analysis delivers the more intuitive result, according to which all board members who voted in favour of continuing with sales are causally, and legally, responsible for the health issues. Let $v_1, \ldots, v_{11}$ designate the individual votes of the board members. Suppose v_1 and v_2 are the votes for stopping sales and production, while the others were in favour of continuing. Let e be the harm that has been done to the health of customers by the product. Then each member of the set $\{v_3, \ldots, v_{11}\}$ satisfies Definition 5 with respect to e. That is, each member who voted in favour of ignoring the early reports is causally, and legally, responsible for the harm to the health of the customers. Take for example v_3. Notice that

the set $C = \{v_3, v_4, v_5, v_6\}$ satisfies condition (R1). For, had the four board members 3, 4, 5, and 6 voted in favour of putting sales and production on hold, sales and production would actually have been stopped. Also, condition (R2) is satisfied. Had up to three of the four board members been in favour of stopping sales and production, then – other things being equal – the company would still have continued to sell and to manufacture the product.[5] Hence, each member of the set C is causally relevant.

Notably, there are several sets C such that $v_3 \in C$, and conditions (R1) and (R2) are satisfied. This is not a problem for Definition 5 since it does not exclude that there are several sets C such that all the conditions are satisfied for an action a being causally, and legally, responsible for an event e. (R3) is satisfied by v_3 since each board member was legally obliged to stop sales and production when it received the early reports about toxicity. Hence, v_3 satisfies our definition of causal responsibility relative to the event that the health of the customers has been harmed.[6]

7 The Halpern-Pearl Approach to Actual Causation

Judea Pearl's seminal *Causality: Models, Reasoning and Inference* (2000) has become the standard reference for causal models. Such models have been used to study both probabilistic and deterministic causation. In his *Actual Causality*, Halpern (2016, Ch. 2) distinguishes between three definitions of actual causation (all of which emerged out of a collaboration with Pearl): (i) the original Halpern-Pearl definition, (ii) the updated Halpern-Pearl definition, and (iii) the modified Halpern-Pearl definition. The Halpern-Pearl approach to token-level causation has been highly influential.

Why did we not build on the Halpern-Pearl approach in this investigation? Our main reason is simplicity. The updated Halpern-Pearl definition is certainly powerful, and allows us to solve overdetermination and related problems. But it does not score well in the dimensions of simplicity and intuitive accessibility. As noted by Liepiņa et al. (2020), the updated Halpern-Pearl definition exceeds the comfort level of most scholars and practitioners of law. The mere framework of causal models, even when confined to deterministic scenarios, is quite demanding. The

[5] Recall that the qualification *other things being equal* is built into the semantics of counterfactuals by some ordering of similarity or relevance.

[6] Again, it is worth noting that collective actions may also be captured along the lines of the NESS test, as shown by Braham and van Hees (2018).

modified Halpern-Pearl definition, first presented in Halpern (2015), is simpler and more accessible, but fails to capture individual causes in scenarios of symmetric overdetermination.[7]

Let us take a closer look at the updated Halpern-Pearl definition to see its vast complexity, even if we present it in a simplified form. We simplify it by assuming that all variables of the respective causal model are binary. This assumption is equivalent to assuming that the causal scenario is modelled using only propositional variables. We further simplify it by leaving implicit the semantics of counterfactual conditionals. Halpern and Pearl use interventionist conditionals instead of the possible-worlds semantics by Lewis (1973b). To avoid confusion at the level of symbolic notation, we use $>$ as symbol for a conditional whose semantics remains unspecified. It may be interpreted as an interventionist conditional, a variably strict conditional in the sense of Lewis (1973b), or a Ramsey Test conditional along the lines of Gärdenfors (1988, Ch. 7).

With these simplifying assumptions in place, we can explain the updated Halpern-Pearl definition as follows. Let X be a set of events, represented by propositional formulas. ϕ is a propositional formula. Let V be the set of propositional variables in terms of which the causal scenario in question is described. X is a cause of an effect ϕ – in a given causal scenario – iff the following conditions hold (cf. Halpern and Pearl (2005, p. 853)):

(1) All members of X and ϕ are true in the actual causal scenario.

(2) There is a partition of the set V into sets W and Z such that

   (a) there is an assignment ν of values to the members of W (which may deviate from the values in the actual causal scenario) in the context of which it holds that $\neg \bigvee X > \neg\phi$, and

   (b) for all $W' \subseteq W$ and all $Z' \subseteq Z$, if the assignment of values to W' agrees with ν and all members of Z' are set to their original values (i.e., the values of the actual causal scenario), then $\bigwedge X > \phi$.

(3) There is no proper subset X' of X such that conditions (1) and (2) are satisfied.

[7] On the latter definition, the set of overdetermining causes qualifies as a cause, but the members of such a set do not (Andreas and Günther 2021).

We think it is obvious that this definition of causation is more complex than our Definition 5. In particular, the application of condition (2b) is computationally rather complex since we need to test whether $\bigwedge X > \phi$ holds for all combinations of subsets of Z and W. Moreover, we need to find the "right partition" of the set V of variables that must satisfy both, condition (2a) and condition (2b). On top of this, it is rather difficult to give an intuitive motivation for condition (2b). This condition may be motivated by ideas about causation as production: the members of Z form an active causal route from the cause to the effect if X is an actual cause (cf. Halpern and Pearl (2005, p. 853) and Hitchcock (2001)). But this intuition does not translate easily into the constructions of condition (2b).

We do not discuss the original Halpern-Pearl definition of causation here for two reasons. First, the updated Halpern-Pearl definition is considered more advanced by their authors. Second, the original Halpern-Pearl definition is almost as complex as the updated one. The only element added in the updated definition is that $\bigwedge X > \phi$ must hold for all $W' \subseteq W$ (where the assignment of values to W' agrees with ν), not only for all $Z' \subseteq Z$ (where all members of Z' are set to their original values) in the context of the value assignment ν for W. The modified Halpern-Pearl definition is significantly simpler than both the original and the updated Halpern-Pearl definition, but fails to capture individual causes in scenarios of symmetric overdetermination. The discussion of token-level causation in Woodward (2003) is guided by the original Halpern-Pearl definition.

Normative considerations come into play when Halpern and Hitchcock (2015) discuss omissions and certain other intricate scenarios which are not yet captured by the updated Halpern-Pearl definition. In essence, they assume a normality ordering among possible worlds which may include considerations of lawful behaviour. If so, a possible world is said to be more normal than another if legal and ethical norms are satisfied to a higher degree and other things are equal. They add the following clause to condition (2a): all possible worlds used for the evaluation of the conditional $\neg \bigvee X > \neg\phi$ are at least as normal as the actual world. In their framework of causal models enriched by a normality ordering, they are able to represent different views about the causal efficacy of omissions to be found in the literature. In particular, they can model the view that an omission is considered a cause only if it involves some violation of a norm (cf. McGrath (2005)).

There is some convergence between Halpern and Hitchcock (2015), Armgardt (2018), and our Definition 5 as regards the treatment of omissions. In all accounts, some distinction is needed between possible worlds in which the behaviour of agents accords with certain norms and others where it does not. But there are also

important differences. While Armgardt (2018) is exclusively interested in some notion of causal responsibility in legal contexts, Halpern and Hitchcock (2015) aim to give a unified account of causation which captures all types of causal scenarios. To achieve this, they add another clause to an account of causation which is already highly complex. Even though Halpern and Hitchcock (2015) do not build normative considerations directly into the semantics of counterfactuals, they modify the evaluation of a specific type of counterfactual. No such modification is needed in our Definition 5 since this definition is not aimed at giving a unified account of causation. Again, we think that our account of causal responsibility merits consideration because it is simpler and intuitively more easily accessible than the Halpern-Pearl account and its extension in Halpern and Hitchcock (2015).[8]

8 Conclusion

Our investigation started from the following two observations. First, extant solutions to the overdetermination problem within the counterfactual approach are technically quite demanding and not easily accessible at an intuitive level. Second, counterfactual reasoning plays an important role in the context of law for the ascription of causal responsibility. This calls for a simpler solution to the overdetermination problem within the counterfactual approach.

Using ideas about iterated difference-making and normatively ideal worlds, we have shown how causal responsibility may be ascribed in a manner that solves the overdetermination problem. Our solution adds some complexity to the simple but-for test, but we have explained and motivated the suggested refinements of this test at an intuitive level. Given some understanding of counterfactuals and hypothetical scenarios, only very simple set-theoretic notions are needed to make sense of our final analysis. In essence, we need to understand what it means that the members of a set of events are present and what it means that the members of such a set are absent. We think our analysis remains at a level of complexity that scholars and practitioners of law feel comfortable with. Moreover, our analysis accounts for the causal responsibility ascribed to omissions who violate norms or legal duties.

It goes without saying that a number of further problems concerning the notion of causal responsibility remain to be dealt with. Problems of preemption are only

[8] Unlike the Halpern-Pearl definitions of causation, our analysis does not fully solve the problem of preemption. While the genuine cause is counted as a cause, the preempted 'cause' is wrongly counted as a cause as well. We plan to remedy this situation in future work.

partially solved by our analysis. Preventions need to be discussed as well. We will address such problems in our ongoing work on causal responsibility.

References

Andreas, H. and Günther, M. (2021). A Ramsey Test Analysis of Causation for Causal Models. *British Journal for the Philosophy of Science* **72**(2): 587–615.

Andreas, H. and Günther, M. (2021). Iterated Difference-Making. *Manuscript* .

Armgardt, M. (2018). Causation in Law, Overdetermination and Normative Ideal Worlds. In *Natural Arguments: A Tribute to John Woods*, edited by D. Gabbay, L. Magnani, W. Park, and A.-V. Pietarinen, London: College Publications. 699–708.

Beebee, H. (2004). Causing and Nothingness. In *Causation and Counterfactuals*, edited by L. A. Paul, E. J. Hall, and J. Collins, Cambridge, MA, USA: MIT Press. 291–308.

Braham, M. and van Hees, M. (2009). Degrees of Causation. *Erkenntnis* **71**(3): 323–344.

Braham, M. and van Hees, M. (2018). Voids or Fragmentation: Moral Responsibility For Collective Outcomes. *The Economic Journal* **128**(612).

Gärdenfors, P. (1988). *Knowledge in Flux*. Cambridge, Mass.: MIT Press.

Halpern, J. Y. (2000). Axiomatizing Causal Reasoning. *Journal of Artificial Intelligence Research* **12**(1): 317–337.

Halpern, J. Y. (2015). A Modification of the Halpern-Pearl Definition of Causality. *Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)* : 3022–3033.

Halpern, J. Y. (2016). *Actual Causality*. Cambridge, MA: MIT Press.

Halpern, J. Y. and Hitchcock, C. (2015). Graded Causation and Defaults. *British Journal for the Philosophy of Science* **66**(2): 413–457.

Halpern, J. Y. and Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science* **56**(4): 843–887.

Hanau, P. (1971). *Die Kausalität der Pflichtwidrigkeit. Eine Studie zum Problem des pflichtmäßigen Alternativverhaltens im bürgerlichen Recht.* Göttingen: Otto Schwartz.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy* **98**(6): 273–299.

Lewis, D. (1973a). Causation. *Journal of Philosophy* **70**(17): 556–567.

Lewis, D. (1973b). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (2000). Causation as Influence. *Journal of Philosophy* **97**(4): 182–197.

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

Liepiņa, R., Sartor, G., and Wyner, A. (2020). Arguing about causes in law: a semi-formal framework for causal arguments. *Artificial Intelligence and Law* **28**(1): 69–89.

McGrath, S. (2005). Causation by Omission: A Dilemma. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* **123**(1/2): 125–148.

Moore, M. S. (2009). *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford University Press.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press, 1st edn.

Schaffer, J. (2010). Contrastive Causation in the Law. *Legal Theory* **16**(4): 259–297.

Woods, J. (2018). Response to Matthias Armgardt. In *Natural Arguments: A tribute to John Woods*, edited by D. Gabbay, L. Magnani, W. Park, and A.-V. Pietarinen, London: College Publications. 709–714.

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