COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 37 (2013) 986–1010
Copyright © 2013 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12059

Compact Representations of Extended Causal Models

Joseph Y. Halpern,[a*] Christopher Hitchcock[b]

[a]*Computer Science Department, Cornell University*
[b]*Division of the Humanities and Social Sciences, California Institute of Technology*

Received 1 February 2012; received in revised form 20 January 2013; accepted 29 January 2013

Abstract

Judea Pearl (2000) was the first to propose a definition of *actual causation* using causal models. A number of authors have suggested that an adequate account of actual causation must appeal not only to causal structure but also to considerations of *normality*. In Halpern and Hitchcock (2011), we offer a definition of actual causation using *extended causal models*, which include information about both causal structure and normality. Extended causal models are potentially very complex. In this study, we show how it is possible to achieve a compact representation of extended causal models.

Keywords: Causality; Normal; Typical; Bayesian network; Compact representation

1. Introduction

One of Judea Pearl's many, many important contributions to the study of causality was the first attempt to use the mathematical tools of causal modeling to give an account of "actual causation," a notion that has been of considerable interest among philosophers and legal theorists (Pearl, 2000, Chapter 10). Pearl later revised his account of actual causation in joint work with Halpern (Halpern & Pearl, 2005). A number of authors (Hall, 2007; Halpern, 2008; Hitchcock, 2007; Menzies, 2004) have suggested that an account of actual causation must be sensitive to considerations of normality, as well as to causal structure. In Halpern and Hitchcock (2011), we suggest a way of incorporating considerations of normality into the Halpern–Pearl theory, and we show how to extend the account to illuminate features of the psychology of causal judgment, as well as features of causal reasoning in the law. Our account of actual causation makes use of "extended causal models," which include both structural equations among a set of variables, and a

Correspondence should be sent to Joseph Y. Halpern, Computer Science Department, 4130 Upson Hall Cornell University, Ithaca, NY 14853. E-mail: halpern@cs.cornell.edu

[*]Supported in part by NSF grants IIS-0812045 and IIS-0911036, AFOSR grant FA9550-08-1-0438 and AFOSR grants FA9550-08-1-0438 and FA9550-05-1-0055, and ARO grant W911NF-09-1-0281.

partial preorder on possible worlds, which represents the relative "normality" of those worlds.

We actually want to think of people as working with the structural equations and normality order to evaluate actual causation. However, consideration of even simple examples immediately suggests a problem. A direct representation of the equations and normality order is too cumbersome for cognitively limited agents to use effectively. If our account of actual causation is to be at all realistic as a model of human causal judgment, some form of compact representation will be needed.

To understand the problem, consider a doctor trying to deal with a patient who has just come in reporting bad headaches. Let us keep things simple, and suppose that the doctor considers only a small number of variables that might be relevant: stress, constriction of blood vessels in the brain, aspirin consumption, and trauma to the head. Again, keeping things simple, assume that each of these variables (including headaches) is binary; that is, it has only two possible values. So, for example, the patient either has a headache or not. Each variable may depend on the value of the other four. To represent that structural equation for the variable "headaches," a causal model will need to assign a value to "headache" for each of the 16 possible values of the other four variables. That means that there are 2^{16}—over 60,000!—possible equations for "headaches." Considering all five variables, there are 2^{80} (over 10^{24}) possible sets of equations. Representing one of these would require eighty binary bits of information. Now consider the normality orders. With five binary variables, there are $2^5 = 32$ possible assignments of values to these variables. Think of each of these assignments as a "possible world." There are 32! (roughly 2.6×10^{35}) strict orders of these 32 worlds, and many more if we allow for ties or incomparable worlds. Altogether, the doctor would need to store close to 200 bits of information just to represent this simple extended causal model.

Now suppose we consider a more realistic model with 50 random variables. Then the same arguments show that we would need as many as $2^{50 \times 2^{49}}$ possible sets of equations, 2^{50} possible worlds, and over $2^{50 \times 2^{50}}$ normality orders (in general, with n binary variables, there are $2^{n2^{n-1}}$ sets of equations, 2^n possible worlds, and $(2^n)! \sim 2^{n2^n}$ strict orders). Thus, with 50 variables, roughly 50×2^{50} bits would be needed to represent a causal model. This is clearly cognitively unrealistic.

The goal of this study is to show that, in practice, representing the information needed to evaluate actual causation can be done in a reasonably compact way, so that the assumption that people are actually doing this is indeed psychologically plausible. To do this, we will make significant use of another of Pearl's signal contributions: the use of directed graphs—specifically, *Bayesian networks*—to represent independence relations (Pearl, 1988).

The first step toward the goal of getting a compact representation comes from the observation that similar representational difficulties arise when it comes to reasoning about probability. For example, if the doctor would like to reason probabilistically about the symptoms, just describing a probability distribution on the 2^{50} worlds would also require 2^{50} (or, more precisely, $2^{50} - 1$) numbers. Bayesian networks allow us to typically get much more compact representations of probability distributions by taking advantage of (conditional) independencies.

Our goal is to arrive at an analog of a Bayesian network representation for both the structural equations and for normality. For the structural equations, it is easy to see where independence comes in. If the equation for each variable X depends on the values of only a few other variables, then the structural equations become much easier to represent. Normality is not typically represented using probability; in Halpern and Hitchcock (2011), we represented it using a partial preorder; in Halpern (2008), Halpern and Pearl (2005), and Huber (2011), it is represented using a *ranking function* (Spohn, 1988). Both of these approaches are instances of what has been called a *plausibility measure* (Friedman & Halpern, 1995). Halpern (2001) has given conditions under which plausibility measures can be represented using Bayesian networks; we apply these ideas here. This allows us to take advantage of conditional independencies to get a compact representation of the normality order, although it is not described probabilistically.

We believe that even greater representational economy may often be possible. This is because we expect that the normality order will often be largely determined by the causal structure. For example, suppose that the causal structure is such that if the patient suffers a head trauma, then he would also suffer from headaches. Then we would expect any world where trauma = 1 and headaches = 1 to be more normal, all else being equal, than a world in which trauma = 1 and headaches = 0. In this way, a representation of causal structure can do "double duty" by representing much of the normality order as well.

An obvious question is whether the normality order induced by the causal structure accurately represents normality. This may not always be the case. In Section 5, we discuss some examples where we might want to have a normality order that does not conform to causal structure in this way. Nevertheless, we would expect that the normality order is largely determined by the equations. Thus, we can get a more compact representation of the normality order by just listing the *exceptions* to the order generated by the equations.

Interestingly, Huber (2011) has suggested an alternative approach to representing causality and normality; rather than using the causal structure to (largely) determine the normality order, we can use the normality order to determine the causal structure. We discuss this possibility in more detail in Section 5.

The rest of this article is organized as follows: In Section 2, we review the basic definitions needed to understand (extended) causal models. In Section 3, we discuss how compact representations of extended causal models can be obtained; this is the technical core of the study. In Section 4, we discuss how we can (typically) get a yet more compact representation by assuming that, by default, it is typical for the variables to obey the structural equations. Finally, in Section 5, we discuss Huber's proposal of using the normality order to represent causality.

2. Extended causal models

Our motivation for extending causal models to incorporate a notion of normality is to address some difficulties facing the Halpern–Pearl definition of actual cause (Halpern &

Pearl, 2005) and to extend it in various ways. We develop the extended account in detail in Halpern and Hitchcock (2011). The Halpern–Pearl approach has been criticized (as have all other approaches to causality!). It is beyond the scope of this study to defend it. In fact, as we shall see, nothing in this study depends on the details of the Halpern–Pearl definition. Our approach to compactly represent extended causal models can be applied to any framework that combines causal models with a normality ordering. In particular, it should be applicable to alternative accounts of actual causation such as Hall (2007).

In this section, we briefly review extended causal models. We encourage the reader to consult Halpern and Pearl (2005) and Halpern and Hitchcock (2011) for more details and motivation. Extended causal models are based on causal models, so we start with a review of causal models.

2.1. Causal models

The description of causal models given here is taken from Halpern (2000); it is a formalization of earlier work of Pearl (1995), further developed in Galles and Pearl (1997), Halpern (2000), and Pearl (2000).

The model assumes that the world is described in terms of *random variables* and their values. For example, if we are trying to determine whether a forest fire was caused by lightning or an arsonist, we can take the world to be described by three random variables:

- FF for forest fire, where $FF = 1$ if there is a forest fire and $FF = 0$ otherwise;
- L for lightning, where $L = 1$ if lightning occurred and $L = 0$ otherwise;
- M for match dropped (by arsonist), where $M = 1$ if the arsonist dropped a lit match, and $M = 0$ otherwise.

Similarly, in an election between Mr. B and Mr. G with 11 voters, we can describe the world by 12 random variables, $V_1, \ldots, V_{11}, W$, where $V_i = 0$ if voter i voted for Mr. B and $V_1 = 1$ if voter i voted for Mr. G, for $i = 1,\ldots,11$, $W = 0$ if Mr. B wins, and $W = 1$ if Mr. G wins.

In these two examples, all the random variables are *binary*; that is, they take on only two values. There is no problem allowing a random variable to have more than two possible values. For example, the random variable V_i could be either 0, 1, or 2, where $V_i = 2$ if i does not vote; similarly, we could take $W = 2$ if the vote is tied, so neither candidate wins.

Some random variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. For example, if we want to model the fact that if the arsonist drops a match *or* lightning strikes then a fire starts, we could use the random variables M, FF, and L as above, with the equation $FF = \max(L, M)$; that is, the value of the random variable FF is the maximum of the values of the random variables M and L. This equation says, among other things, that if $M = 0$ and $L = 1$, then $FF = 1$. The equality sign in this equation should be thought of more like an assignment statement in programming languages; once we set the values of M and L, then the value of FF is set to their maximum. However, despite the equality, if a forest fire starts some other way, that does not force the value of either M or L to be 1. That is, setting FF to 1 does not result in either M or L being set to 1.

Alternatively, if we want to model the fact that a fire requires both a lightning strike *and* a dropped match (perhaps the wood is so wet that it needs two sources of fire to get going), then the only change in the model is that the equation for FF becomes $FF = \min(L, M)$; the value of FF is the minimum of the values of M and L. The only way that $FF = 1$ is if both $L = 1$ and $M = 1$.

It is conceptually useful to split the random variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model; and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. In the forest fire example, the variables M, L, and FF are endogenous. We do not want to concern ourselves with the factors that make the arsonist drop the match or the factors that cause lightning. Thus, we do not include endogenous variables for these factors. Instead, we introduce a single exogenous variable U whose values take the form (i, j), where i and j each take the value 0 or 1. The value of U will then determine the values of M and L.[1]

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *modifiable structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $\mathcal{Y} \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y *ranges*). As suggested above, in the forest fire example, we have $\mathcal{U} = \{U\}$, where U is the exogenous variable, $\mathcal{R}(U)$ consists of the four possible values of U discussed earlier, $\mathcal{V} = \{FF, L, M\}$, and $\mathcal{R}(FF) = \mathcal{R}(L) = \mathcal{R}(M) = \{0, 1\}$.

$\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$, a function denoted F_X such that

$$F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X).$$

This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. If there is one exogenous variable U and three endogenous variables, X, Y, and Z, then F_X defines the values of X in terms of the values of Y, Z, and U. For example, we might have $F_X(u, y, z) = u + y$, which is usually written as $X = U + Y$.[2] Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.

In the running forest fire example, where U has four possible values of the form (i,j), the i value determines the value of L and the j value determines the value of M. Although F_L gets as arguments the values of U, M, and FF, in fact, it depends only on the (first component of the) value of U; that is, $F_L((i,j), m, f) = i$. Similarly, $F_M((i,j), l, f) = j$. In this model, the value of FF depends only on the value of L and M. *How* it depends on them depends on whether we are considering the conjunctive model or the disjunctive model.

A *possible world* is an assignment of values to all the endogenous random variables in a causal model. We might use the term "small world" to describe such an assignment, to distinguish it from a "large world," which is an assignment of values to all the variables in a causal model, both endogenous and exogenous.

In other applications, we may well need to make use of such "large worlds." In this paper, however, we need to use only small worlds.[3] Hence, in this study, "possible world" and "world" should be understood as referring to such "small worlds." Intuitively, a possible world is a maximally specific description of a situation within the language allowed by the set of endogenous variables. Thus, a world in the forest fire example might be one where $M = 1$, $L = 0$, and $FF = 0$; the match is dropped, there is no lightning, and no forest fire. As this example shows, a possible world does not have to satisfy the equations of the causal model.

A causal model M is *acyclic* if its equations are such that there is no sequence of variables $X_1, X_2, \ldots, X_n$, where X_2 depends nontrivially on $X_1, \ldots, X_n$ depends nontrivially on X_{n-1}, and X_1 depends nontrivially on X_n. In an acyclic causal model, a complete specification $\vec{U} = \vec{u}$ of the value(s) of the exogenous variable(s), called a *context*, uniquely determines the values of all the endogenous variables. In other worlds, given an acyclic causal model M, the choice of a context $\vec{u}$ suffices to determine a possible world. If the variable X takes the value x in this possible world, we write $(M, \vec{u}) \models X = x$. In the sequel, we consider only acyclic causal models; these seem to be rich enough to deal with essentially all the examples of causality that have been considered in the literature (c.f., the discussion in Halpern & Pearl, 2005).

Structural equations do more than just constrain the possible values of the variables in a causal model. They also determine what happens in the presence of external *interventions*. For example, we can explain what would happen if one were to intervene to prevent the arsonist from dropping the match. In the disjunctive model, there is a forest fire exactly if there is lightning; in the conjunctive model, there is definitely no fire. An "intervention" does not necessarily imply human agency. The idea is rather that some independent process overrides the existing causal structure to determine the value of one or more variables, regardless of the value of its (or their) usual causes. Woodward (2003) gives a detailed account of such interventions. Setting the value of some endogenous variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ by means of an intervention results in a new causal model denoted $M_{X=x}$. $M_{X=x}$ is identical to M, except that the equation for X in $\mathcal{F}$ is replaced by $X = x$. Given a context $\vec{u}$, $(M_{X=x}, \vec{u})$ may be thought of as the possible world that would result from intervening to set the variable X to the value x. It is this ability to represent the effects of interventions on the system that gives a causal model its distinctively "causal" character.

If there are distinct values x, x' of X, and y, y' of Y such that $(M_{X=x}, \vec{u}) \models Y = y$ and $(M_{X=x'}, \vec{u}) \models Y = y'$, then we say that Y *counterfactually depends* on X in $(M, \vec{u})$. Intuitively, this means that intervening on the value of X can make a difference for the value of Y.

We regard a causal model as representing objective features of the world. More precisely, given a choice of endogenous and exogenous variables, there is a correct choice of functions to represent the causal dependence of the variables upon one another. The correctness of a causal model can be tested, at least in principle, by performing the relevant interventions on the values of the variables.

2.2. Actual causation

One relation that has attracted considerable attention, especially in philosophy and legal theory, is *actual causation*. For example, the claim that the arsonist's lighting his match caused the forest fire describes a relation of actual causation. This claim is expressed after the fact, and it implies that the arsonist did light his match, and that the forest fire occurred. In addition, it asserts that the lit match is among the actual causes of the forest fire. Relations of actual causation cannot simply be "read off" a causal model. Our model of the forest fire tells us what would happen if the arsonist lights his match, and if lightning strikes, but it does not tell us whether either of these events would count as a cause of the fire. Actual causation has been of interest, in part, because it seems to be involved in assessments of moral and legal responsibility.

The full definition of actual causation offered in Halpern and Pearl (2005) is fairly complex, and most of the details do not matter to the present discussion. It suffices to note that according to the Halpern–Pearl definition, counterfactual dependence is *sufficient* for actual causation. More precisely, $X = x$ is an actual cause of $Y = y$ in $(M, \vec{u})$ if (but not only if): (a) $(M, \vec{u}) \models X = x, Y = y$; and (b) there exist $x' \neq x$ and $y' \neq y$ such that $(M_{X=x'}, \vec{u}) \models Y = y'$. The new model $M_{X=x'}$, together with the context $\vec{u}$, determines a unique value for each of the endogenous variables. This assignment of values determines a possible world, which is called a *witness* to $X = x$ being an actual cause of $Y = y$.[4] Counterfactual dependence is not necessary for actual causation, as counterfactual dependence can fail in cases of preemption and overdetermination. But we can ignore these cases for now.

Halpern and Pearl (2005) already noted that a causal model does not suffice to determine causality. There are subtle examples that can be characterized by causal models that are isomorphic, but where the judgment of actual causation differs. One approach to solving these problems, suggested by Halpern and Pearl (2005), and developed in different ways by Hall (2007), Halpern (2008), Hitchcock (2007), and Halpern and Hitchcock (2011), is to incorporate considerations about *defaults*, *typicality*, and *normality*. "Normality" and its cognates ("normal," "norm," "abnormal," etc.) tend to be ambiguous. They can refer to statistical frequency, as when we say that there has been more rain than normal for this time of year. But they can also refer to prescriptive rules, as when we say that someone has violated a moral norm. These concepts obviously differ in important ways, but in ordinary thought we often slip between the two ideas without even realizing it. Our conjecture is that these two different senses of "normality" affect causal judgments in roughly similar ways, so we have left the word deliberately ambiguous. (We remark that there are other interpretations of normality as well; see Halpern & Hitchcock, 2010, for further discussion.)

Here is a simple example to illustrate how considerations of normality can affect causal judgments:

Example 2.1

Professor Smith and an administrative assistant take the last two pens in the department office. There is a department rule that administrative assistants are allowed to take the pens, whereas faculty are not. Later, a problem arises from the lack of pens.

Knobe and Fraser (2008) presented subjects with a version of this vignette and asked them to rate their agreement with *either* the statement that Professor Smith caused the problem, or that the administrative assistant caused the problem. Subjects were much more strongly inclined to agree that Professor Smith caused the problem.

We can model this case as follows. (For simplicity, we ignore the exogenous variable (s).) Let $PS = 1$ if Professor Smith takes a pen, $PS = 0$ if not; $AA = 1$ if the administrative assistant takes a pen, $AA = 0$ if not; and $PO = 1$ if the problem occurs, $PO = 0$ if not. Then the equation for PO will be $PO = \min(PS, AA)$. It should be apparent that the dependence of PO on PS and AA is symmetric; in particular, PO counterfactually depends on both variables. Nonetheless, judgments about the two are different. Professor Smith violated a norm, whereas the administrative assistant did not, and this difference seems to be affecting causal judgments about the case.

According to the theory of Halpern and Hitchcock (2011), potential causes are "graded" according to the normality of their witnesses.[5] In the pen vignette, the witness for $PS = 1$ being an actual cause of $PO = 1$ is the world $(PS = 0, AA = 1, PO = 0)$; the witness for $AA = 1$ being an actual cause is $(PS = 1, AA = 0, PO = 0)$. As Professor Smith's taking a pen violates a norm, the former world is more normal, and $PS = 1$ receives a higher causal grading.

This kind of treatment can be extended to a wide range of cases. For example, we can make the familiar distinction between *causes* and *background conditions*. Suppose that an arsonist lit a match, oxygen was present in the air, and a fire occurred. The fire counterfactually depends on both the match and the oxygen, but we tend to consider only the match as the cause of the fire, viewing the oxygen as a mere background condition. By regarding a world with oxygen and no match as more normal than a world with a lit match and no oxygen, we can treat this case in a way that is formally analogous to the treatment of pen vignette. Halpern and Hitchcock (2011) provide a number of further illustrations.

Some will worry that this account of actual causation will make causation subjective. Although we agree that this introduces a subjective element to actual causation, we do not view this as a concern. The causal model represents the objective core of causation. The patterns of causal dependence represented by the equations of a causal model are objective features of the world. *Actual causation* is a further relation that goes beyond these objective dependence relations. It is determined *in part* by objective relations of causal dependence, but it is also determined in part by considerations of normality.

2.3. Extended causal models

Following our earlier work (Halpern & Hitchcock, 2011), we formalize normality using *extended causal models*. We assume that there is a partial preorder $\succeq$ over worlds; $s \succeq s'$ means that world s is at least as normal as world s'. The fact that $\succeq$ is a partial preorder means that it is reflexive (for all worlds s, we have $s \succeq s$, so s is at least as normal as itself) and transitive (if s is at least as normal as s' and s' is at least as normal as s'', then s is at least as normal as s'').[6] We write $s \succ s'$ if $s \succeq s'$ and it is not the case that $s' \succeq s$, and $s \equiv s'$ if $s \succeq s'$ and $s' \succeq s$. Thus, $s \succ s'$ means that s is strictly more normal than s', whereas $s \equiv s'$ means that s and s' are equally normal. Note that we are not assuming that $\succeq$ is total; it is quite possible that there are two worlds s and s' that are incomparable as far as normality. The fact that s and s' are incomparable does *not* mean that s and s' are equally normal. We can interpret it as saying that the agent is not prepared to declare either s or s' as more normal than the other, and also not prepared to say that they are equally normal; they simply cannot be compared in terms of normality. An *extended causal model* is a tuple $M = (\mathcal{S}, \mathcal{F}, \succeq)$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and $\succeq$ is a partial preorder on worlds, which can be used to compare how normal different worlds are.

Partial preorders are essentially used by Kraus, Lehmann, and Magidor (1990) and Shoham (1987) to model normality. Many other approaches to modeling normality have been proposed in the literature, including *ε−semantics* (Adams, 1975; Geffner, 1992; Pearl, 1989), *possibility measures* (Dubois & Prade, 1991), and *ranking functions* (Goldszmidt & Pearl, 1992; Spohn, 1988). Perhaps the most general approach uses what are called *plausibility measures* (Friedman & Halpern, 1995, 2001); we return to plausibility measures below. Some of these approaches (specifically, ε-semantics, possibilistic structures, and ranking functions) essentially impose a total order on worlds; as we shall see, the greater generality of partial orders provides a useful modeling tool. That said, almost all of what we say in this article applies to all these other approaches as well.

3. Compact representations of extended models

In Halpern and Hitchcock (2011), a formal definition of actual causality is given; the definition is given relative to an extended causal model. To determine actual causation according to this definition, an agent would have to have a representation of the model. As we suggested in the introduction, a naïve representation of a model involving n binary random variables would involve $n2^{n-1}$ values, as for each variable X_i, the function F_{X_i} has to give the value of X_i for each of the 2^{n-1} settings of the other variables. Even if we restrict attention to acyclic models, there may be one variable X that depends on all the others, so that the function F_X corresponding to X has to give a value to X for each of the 2^{n-1} settings of the other variables. Moreover, we must still define a partial preorder of the 2^n worlds. Even if we restrict to total orders or use one of the other representations of normality, as there are $2^n! \sim 2^{n2^n}$ total orders of the worlds, this requires at least $n2^n$

bits of information. Nevertheless, as we now show, in practice, it will often be possible to represent this information in a far more compact way.

3.1. Representing causal equations compactly: The big picture

As we mentioned in the introduction, the key tool for getting compact representations is the use of graphical representations. It is sometimes helpful to represent a causal model graphically. Each node in the graph corresponds to one variable in the model. An arrow from one node, say L, to another, say FF, indicates that the former variable figures as a nontrivial argument in the equation for the latter—that is, the latter depends on the former. Thus, we could represent either the conjunctive or the disjunctive model of the forest fire example using Fig. 1. Note that the graph conveys only the qualitative pattern of dependence; it does not tell us how one variable depends on others. Thus, the graph alone does not allow us to distinguish between the disjunctive and the conjunctive models.

The semantics (i.e., meaning) of such a graphical representation depends on how it is being used. In the case of causal models, it is particularly simple. The value of a variable in a graph depends only on the values of its parents and is independent of the values of all other variables once the values of its parents are given. Thus, in the forest fire example, when we write the equation for FF, it depends only on the value of L and M, and not on the value of U. (Of course, the values of L and M depend on the value of U, so indirectly the value of FF depends on the value of U.) Formally, this means that F_{FF} needs to take only two arguments (the value of L and the value of M), rather than three. More generally, if each variable in a graph with n nodes has at most k parents, we can describe the equations using $n2^k$ values. If k is small relative to n (as it is in many cases), this can be considerably less than $n2^{n-1}$, and thus be computationally feasible.

It is not always the case that all nodes have few parents. Consider the voting example. In that case, the outcome depends on how all the voters vote; that is, all of $V_1, \ldots, V_{11}$ are parents of W. Thus, to describe how W depends on the other variables, we must

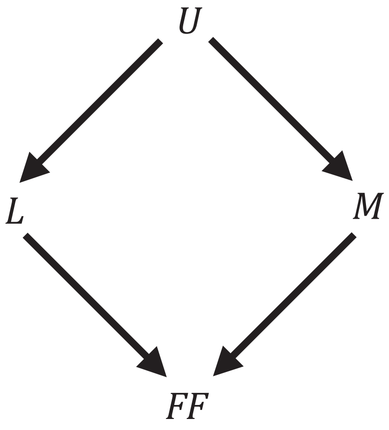

Fig. 1. A graphical representation of structural equations.

specify the outcome for each of the 2^{11} ways that the voters could vote. But we can do this simply without needing to list 2^{11} separate values: We simply say that $W = 1$ if $V_1 + \cdots + V_{11} \geq 6$. In the forest fire example, we can replace the explicit description of the value of FF in terms of the four possible settings of L and M by just writing $FF = \max(L, M)$ or $FF = \min(L, M)$, depending on whether we are considering the disjunctive or conjunctive model. There are times when the best we can do is to write out the explicit description of a function. But in most cases of interest, there will be a much more compact description. The bottom line here is we expect to be able to represent the structural equations compactly in most cases of interest.

3.2. Representing the normality relation compactly: The big picture

We can also make use of a graphical representation to represent the normality order compactly. It is well known that *Bayesian networks* can provide a compact representation for a probability distribution. Suppose that we have a set $V = \{X_1, X_2, \ldots, X_n\}$ of variables, the worlds determined by these variables are those of the form $(x_1, \ldots, x_n)$, where x_i is a possible value of X_i. If we have n binary variables, and thus 2^n possible worlds, we need 2^n numbers to describe a probability distribution on these worlds. (Actually, we need only 2^{n-1}, as the sum of the numbers must be 1.) A *quantitative Bayesian network* on a set $V = \{X_1, \ldots, X_n\}$ is an ordered pair (G, f), where G is a *Bayesian network*, that is, a directed acyclic graph, with n nodes, each labeled by a different variable in V and f associates with each variable X_i in V a *conditional probability table (cpt) for*X_i, which describes the probability that $X_i = 1$ conditional on all the possible settings of its parents in G. For example, if the parents of X_i are X_j, X_k, and X_l, then we must know the probability that $X_i = 1$ conditional on each of the eight possible settings of X_j, X_k, and X_l. A probability measure Pr on the worlds determined by the variables in V is *represented* by (G, f) if (a) G satisfies the *Markov condition*, namely, that each variable X_i is independent of its nondescendants conditional on its parents;[7] and (b) the cpt $f(X_i)$ correctly describes the probability (according to Pr) of X_i conditional on all possible values of its parents. If (G, f) represents Pr, then we can recover Pr from (G, f) using quite straightforward computations (see Halpern, 2003; Pearl, 1988).

Given a probability distribution Pr, it is always possible to find a quantitative Bayesian network (G, f) that represents it. Moreover, if each node in G has at most k parents, then we need at most $(2^k - 1)n$ numbers to describe (G, f), as each cpt requires at most $2^k - 1$ numbers.

Our representation of normality does not use probabilities; rather, it uses a partial preorder on worlds. However, as pointed out by Halpern (2001), the "technology" of Bayesian networks can be applied to mathematical structures other than probability. We just need a structure that has a number of minimal properties and has an analog of (conditional) independence (so that we can have an analog of the Markov condition). Results of Friedman and Halpern (2001) show that a partial order on worlds gives us just such a structure. Moreover, other representations of normality that have been considered in the literature (e.g., ranking functions and possibility measures) also have such a structure, so can also be represented using Bayesian networks. In the language of Friedman and Halpern (1995, 2001), a sufficient condition for a representation of uncertainty to be represented using Bayesian network is that it can be viewed as an *algebraic conditional plausibility measure*. We briefly review some of the relevant details here.

Plausibility measures, introduced by Friedman and Halpern (1995, 2001), are intended to be a generalization of all standard approaches in representing uncertainty. The basic idea behind plausibility measures is straightforward. A probability measure on a finite set W of worlds maps subsets of W to [0,1]. A *plausibility measure* is more general; it maps subsets of W to some arbitrary set D partially ordered by $\leq$. If Pl is a plausibility measure, $\mathrm{Pl}(U)$ denotes the plausibility of U. If $\mathrm{Pl}(U) \leq \mathrm{Pl}(V)$, then V is at least as plausible as U. Because the order is partial, it could be that the plausibility of two different sets is incomparable. An agent may not be prepared to order two sets in terms of plausibility. D is assumed to contain two special elements, $\perp$ and $\top$, such that $\perp \leq d \leq \top$ for all $d \in D$. We require that $\mathrm{Pl}(\emptyset) = \perp$ and $\mathrm{Pl}(W) = \top$. Thus, $\perp$ and $\top$ are the analogs of 0 and 1 for probability. We further require that if $U \subseteq V$, then $\mathrm{Pl}(U) \leq \mathrm{Pl}(V)$. This seems reasonable; a superset of U should be at least as plausible as U. As Bayesian networks make such heavy use of conditioning, we need to deal with *conditional* plausibility measures (cpms), not just plausibility measures. A conditional plausibility measure maps pairs of subsets of W to some partially ordered set D. We write $\mathrm{Pl}(U \mid V)$ rather than $\mathrm{Pl}(U,V)$ in keeping with standard notation for conditioning. We typically write just $\mathrm{Pl}(U)$ rather than $\mathrm{Pl}(U \mid W)$ (so unconditional plausibility is identified with conditioning on the whole space).

In the case of a probability measure Pr, it is standard to take $\Pr(U \mid V)$ to be undefined if $\Pr(V) = 0$. In general, we must make precise what the allowable second arguments of a cpm are. For simplicity here, we assume that $\mathrm{Pl}(U \mid V)$ is defined as long as $V \neq \emptyset$. For each fixed $V \neq \emptyset$, $\mathrm{Pl}(\cdot \mid V)$ is required to be a plausibility measure on W. More generally, we require the following properties:

CP 11. $\mathrm{Pl}(\emptyset \mid V) = \perp$.
CP 12. $\mathrm{Pl}(W \mid V) = \top$.
CP 13. If $U \subseteq U'$, then $\mathrm{Pl}(U \mid V) \leq \mathrm{Pl}(U' \mid V)$.
CP 14. $\mathrm{Pl}(U \mid V) = \mathrm{Pl}(U \cap V \mid V)$.

Conditional probability satisfies additional properties; for example, $\Pr(U \mid V') = \Pr(U \mid V) \times \Pr(V \mid V')$ if $U \subseteq V \subseteq V'$ and $V \neq \emptyset$, and $\Pr(V_1 \cup V_2 \mid V) = \Pr(V_1 \mid V) + \Pr(V_2 \mid V)$ if V_1 and V_2 are disjoint sets. These properties turn out to play a critical role in carrying out the reasoning used in Bayesian networks. We want plausibilistic analogs of them. This requires us to have plausibilistic analogs of addition and multiplication, so that we can take $\mathrm{Pl}(V_1 \cup V_2 \mid V_3) = \mathrm{Pl}(V_1 \mid V_3) \oplus \mathrm{Pl}(V_2 \mid V_3)$ if V_1 and V_2 are disjoint, and $\mathrm{Pl}(V_1 \mid V_3) = \mathrm{Pl}(V_1 \mid V_2) \otimes Pl(V_2 \mid V_3)$ if $V_1 \subseteq V_2 \subseteq V_3$. We give the formal definitions in the next section. For now, we just note that these properties hold for probability if we take $\oplus$ and $\otimes$ to be + and ×, respectively; and they hold for ranking functions if we take $\oplus$ and $\otimes$ to be min and +. They also hold for possibility measures, although the situation is somewhat more complicated there (see Halpern, 2001).

We want to view a partial preorder on worlds as an algebraic conditional plausibility measure. There is a problem though. A plausibility measure attaches a plausibility to *sets* of worlds, not single worlds. So, given a partial preorder $\succcurlyeq$ on worlds, we must first find a plausibility measure $\mathrm{Pl}_{\succeq}$ with the property that $\mathrm{Pl}_{\succeq}(w_1) \geq \mathrm{Pl}_{\succeq}(w_2)$ if $w_1 \succeq w_2$. We then need to show how we can extend this plausibility measure to an algebraic conditional plausibility measure. We do this in the next section.

We are mainly interested in algebraic plausibility measures on a set of worlds determined by random variables $X_1, \ldots, X_n$. By results of Halpern (2001), such plausibility measures can be represented using a *plausibilistic* Bayesian network (G, f), where G is a Bayesian network and f associates with each node X_i in G a *conditional plausibility table*; we abuse notation and use the abbreviation cpt for a conditional plausibility table as well. The cpt for X specifies the plausibility of each possible value of X, conditional on all possible values of X's parents. (Note that if X is binary, it no longer suffices to just specify the plausibility of $X = 1$ conditional on X's parents because the plausibility of $X = 1$ does not necessarily determine the plausibility of $X = 0$ conditional on X's parents.) We can define what it means for a plausibilistic Bayesian network to represent a plausibility measure just as in the probabilistic case. And, just as in the probabilistic case, a plausibility measure that takes values in an algebraic cpm can be represented by plausibilistic Bayesian network. Moreover, if each node in the network has relatively few parents, we have a compact representation of the plausibility measure. The bottom line here is that, once we show how to represent a partial preorder as an algebraic plausibility measure, we can get a representation of the preorder using Bayesian networks that will typically be compact.

In addition to representing the Bayesian network, if we use a plausibility measure, we must also represent the plausibility domain; that is, we have to describe its elements and the ordering on them. We expect that, typically, the domain will be relatively small, and the ordering easy to describe. Indeed, here the fact that we allow partial orders makes it easier because we allow many elements to be incomparable. This will become clearer in the examples in Section 4.

3.3. Representing the normality relation compactly: The technical details

In this section, we fill in the technical details for the results discussed in the previous section. This section can be skipped without loss of continuity. We start with the formal definition of algebraic conditional plausibility measures.

Definition 3.1

An *algebraic conditional plausibility measure* Pl on W maps pairs of subsets of W to a domain D that is endowed with operations $\oplus$ and $\otimes$, defined on domains $Dom(\oplus)$ and $Dom(\otimes)$, respectively, such that the following properties hold:

Alg1. If V_1 and V_2 are disjoint subsets of W and $V \neq \emptyset$, then $\mathrm{Pl}(V_1 \cup V_2 \mid V) = \mathrm{Pl}(V_1 \mid V) \oplus \mathrm{Pl}(V_2 \mid V)$.

Alg2. If $U \subseteq V \subseteq V'$ and $V \neq \emptyset$, then $\mathrm{Pl}(U \mid V') = \mathrm{Pl}(U \mid V) \otimes \mathrm{Pl}(V \mid V')$.

Alg3. $\otimes$ distributes over $\oplus$; more precisely, $a \otimes (b_1 \oplus \cdots \oplus b_n) = (a \otimes b_1) \oplus \cdots \oplus (a \otimes b_n)$ if $(a, b_1), \ldots, (a, b_n), (a, b_1 \oplus \cdots \oplus b_n) \in Dom(\otimes)$ and $(b_1, \ldots, b_n)$, $(a \otimes b_1, \ldots, a \otimes b_n) \in Dom(\oplus)$, where $Dom(\oplus) = \{(\mathrm{P1}(V_1 \mid U), \ldots, \mathrm{P1}(V_n \mid U)) : V_1, \ldots, V_n$ are pairwise disjoint and $U \neq \emptyset\}$, and $Dom(\otimes) = \{(\mathrm{P1}(U \mid V), \mathrm{P1}(V \mid V')) : U \subseteq V \subseteq V', V \neq \emptyset\}$. (The reason that this property is required only for tuples in *Dom*(⊕) and *Dom*(⊗) is discussed shortly. Note that parentheses are not required in the expression $b_1 \oplus \cdots \oplus b_n$, although, in general, ⊕ need not be associative. This is because it follows immediately from Alg1 that ⊕ is associative and commutative on tuples in *Dom*(⊕).)

Alg4. If $(a,c),(b,c) \in Dom(\otimes)$, $a \otimes c \leq b \otimes c$, and $c \neq \perp$, then $a \leq b$.

The restrictions in Alg3 and Alg4 to tuples in *Dom*(⊕) and *Dom*(⊗) make these conditions a little more awkward to state. It may seem more natural to consider a stronger version of, say, Alg4 that applies to all pairs in $D \times D$. Roughly speaking, *Dom*(⊕) and *Dom* (⊗) are the only tuples where we really care how ⊕ and ⊗ work. We use ⊕ to determine the (conditional) plausibility of the union of two disjoint sets. Thus, we care about $\mathrm{P1}(V_1 \mid V)$ and $\mathrm{P1}(V_2 \mid V)$, respectively, where V_1 and V_2 are disjoint sets, in which case we want $a \oplus b$ to be $\mathrm{P1}(V_1 \cup V_2 \mid V)$. More generally, we care about $a_1 \oplus \cdots \oplus a_n$ only if a_i has the form $\mathrm{P1}(V_i \mid V)$, where $V_1, \ldots, V_n$ are pairwise disjoint. *Dom*(⊕) consists of precisely these tuples of plausibility values. Similarly, we care about $a \otimes b$ only if a and b have the form $\mathrm{P1}(U_1 \mid U_2)$ and $\mathrm{P1}(U_2 \mid U_3)$, respectively, where $U_1 \subseteq U_2 \subseteq U_3$, in which case we want $a \otimes b$ to be $\mathrm{P1}(U_1 \mid U_3)$. *Dom*(⊗) consists of precisely these pairs (a,b). By requiring that Alg3 and Alg4 hold only for tuples in *Dom*(⊕) and *Dom*(⊗) rather than on all tuples in $D \times D$, some cpms of interest become algebraic that would otherwise not be. (see Halpern, 2001, 2003 for examples.) Restricting ⊕ and ⊗ to *Dom*(⊕) and *Dom*(⊗) will also make it easier for us to view a partial preorder as an algebraic plausibility measure. As ⊕ and ⊗ are significant mainly to the extent that Alg1 and Alg2 hold, and Alg1 and Alg2 apply to tuples in *Dom*(⊕) and *Dom*(⊗), respectively, it does not seem unreasonable that properties like Alg3 and Alg4 be required to hold only for these tuples.

In an algebraic cpm, we can define a set U to be *plausibilistically independent of V conditional on V'* if $V \cap V' \neq \emptyset$ implies that $\mathrm{P1}(U \mid V \cap V') = \mathrm{P1}(U \mid V')$. The intuition here is that learning V does not affect the conditional plausibility of U given V'. Note that conditional independence is, in general, asymmetric. U can be conditionally independent of V without V being conditionally independent of U. Although this may not look like the standard definition of probabilistic conditional independence, it is not hard to show that this definition agrees with the standard definition (that $\Pr(U \cap V \mid V') = \Pr(U \mid V') \times \Pr(V \mid V')$) in the special case that the plausibility measure is actually a probability measure (see Halpern, 2001 for further discussion of this issue). Of course, in this case, the definition is symmetric.

The next step is to show how to represent a partial preorder on worlds as an algebraic plausibility measure. We do so using ideas of Friedman and Halpern (2001).

Suppose that we have an extended causal model $M = (\mathcal{S}, \mathcal{F}, \succeq)$. We proceed as follows. Define a preorder $\succeq^+$ on subsets of W by taking $U \succeq^+ V$ if, for all $w \in V$, there exists some $w' \in U$ such that $w' \succeq w$. It is easy to check that $w \succeq w'$ if $\{w\} \succeq^+ \{w'\}$. Thus, we have a partial preorder on sets that extends the partial preorder on worlds. We might consider getting an unconditional plausibility measure $\mathrm{P1}_{\succeq}$ that extends the partial preorder on worlds by taking the range of $\mathrm{P1}_{\succeq}$ to be subsets of W, and defining $\mathrm{P1}_{\succeq}$ as the identity; that is, taking $\mathrm{P1}_{\succeq}(U) = U$, and taking $U \geq V$ if $U \succeq V$.

This almost works. There is a subtle problem though. The relation $\geq$ used in plausibility measures must be an order, not a preorder. For an order $\geq$ on a set X, if $x \geq x'$ and $x' \geq x$, we must have $x' = x$. Thus, for example, if $w \succeq w'$ and $w' \succeq w$, then we want $\mathrm{P1}_{\succeq}(\{w\}) = \mathrm{P1}_{\succeq}(\{w'\})$. This is easily arranged.

Define $U \equiv V$ if $U \succeq^+ V$ and $V \succeq^+ U$. Let $[U] = \{U' \subseteq W : U' \equiv U\}$. Now if we take $\mathrm{P1}_{\succeq}(U) = [U]$, and take $[U] \geq [V]$ if $U' \succeq V'$ for some $U' \in [U]$ and $V' \in [V]$, then it is easy to check that $\geq$ is well defined (as if $U' \succeq V'$ for some $U' \in [U]$ and $V' \in [V]$, then $U' \succeq V'$ for all $U' \in [U]$ and $V' \in [V]$) and is a partial order.

Although this gives us an unconditional plausibility measure extending $\succeq$, we are not quite there yet. We need a conditional plausibility measure, and a definition of $\oplus$ and $\otimes$. Note that if $U_1, U_2 \in [U]$, then $U_1 \cup U_2 \in [U]$. As W is finite, it follows that each set $[U]$ has a largest element, namely, the union of the sets in $[U]$.

Let D be the domain consisting of $\perp$, $\top$, and all elements of the form $d_{[U]|[V]}$ for all $[U]$ and $[V]$ such that the largest element in $[U]$ is a strict subset of the largest element in $[V]$. We place an ordering $\geq$ on D by taking $\perp < d_{[U]|[V]} < \top$ and $d_{[U]|[V]} \leq d_{[U']|[V']}$ if $[V] = [V']$ and $U' \succeq^+ U$. We view D as the range of an algebraic plausibility measure, defined by taking

$$\mathrm{P1}_{\succeq}(U \mid V) = \begin{cases} \perp & \text{if } U \cap V = \emptyset \\ \top & \text{if } U \cap V = V \\ d_{[U \cap V]|[V]} & \text{otherwise.} \end{cases}$$

We can define $\oplus$ and $\otimes$ on D so that Alg1 and Alg2 hold. This is easy to do, in large part because we only need to define $\oplus$ and $\otimes$ on *Dom*($\oplus$) and *Dom*($\otimes$), where the definitions are immediate because of the need to satisfy Alg1 and Alg2. It is easy to see that these conditions and the definition of P1 guarantee that Pl1–4 hold. With a little more work, it can be shown that these conditions imply Alg3 and Alg4 as well. (Here the fact that Alg3 and Alg4 are restricted to *Dom*($\oplus$) and *Dom*($\otimes$) turns out to be critical; it is also important that $U \equiv V$ implies that $U \equiv U \cup V$.)

This construction gives us an algebraic cpm, so the results of Halpern (2001) apply. Specifically, we can represent $\succeq$ using a Bayesian network (G, f). The structure of G is determined by the independencies exhibited by the normality order on worlds. There is no guarantee that G will be the same as the graph that represents the causal structure. In many cases, however, there will be substantial overlap between the Bayesian network representation of the normality order and the causal model. As we show in Section 4 below, when this occurs, even greater economy is possible.

3.4. Using a compact representation to determine a normality order

The discussion above shows that if we start with an algebraic conditional plausibility measure determined by a normality order on worlds, then we can represent it using a Bayesian network. Moreover, this representation will often be compact. But what we really want is more like the converse. Suppose that we are given a quantitative Bayesian network. Can we use that to determine a normality order on worlds? The reason that we are particularly interested in this question is that, in many cases of interest, it is quite natural to characterize a situation using a quantitative Bayesian network.

Suppose, for example, that a lawyer is arguing that a defendant should be convicted of arson. The lawyer will attempt to establish a claim of actual causation: that the defendant's action of lighting a match was an actual cause of the forest fire. To do this, he or she will need to convince the jury that a certain extended causal model is correct, and that certain initial conditions obtained (e.g., that the defendant did indeed light a match). To justify a causal model, she will need to defend the equations. This might involve convincing the jury that the defendant's match was the sort of thing that could cause a forest fire (the wood was dry), and that there would have been no fire in the absence of some triggering event, such as a lightning strike or an act of arson.

The lawyer will also have to defend a normality ordering. To do this, she might argue that a lightning strike could not have been reasonably foreseen; that lighting a fire in the forest at that time of year was in violation of a statute; and that it was to be expected that a forest fire would result from such an act. The key idea here is that it will usually be easier to justify a normality ordering in a *piecemeal* fashion. Instead of arguing for a particular normality ordering on entire worlds, she argues that individual variables typically take certain values in certain situations.[8] In doing this, she is defending a particular cpt for each variable. What we show in this section is that having a cpt for each variable leads in a natural way to a particular choice of normality ordering on worlds.

Recall that the Bayesian network G is labeled by variables $X_1,\ldots,X_n$; we want to define a normality order on worlds of the form $(x_1,\ldots,x_n)$, where x_i is a possible value of the random variable X_i. We will associate with each such world a plausibility value of the form $a_1 \otimes \cdots \otimes a_n$, where a_i is a value in the cpt for X_i. For example, if $n = 3$, and, according to G, X_1 and X_2 are independent and X_3 depends on both X_1 and X_2, then a world (1,0,1) would be assigned a plausibility of $a_1 \otimes a_2 \otimes a_3$, where a_1 is the unconditional plausibility of $X_1 = 1$ according to the cpt for X_1, a_2 is the unconditional plausibility of $X_2 = 0$ according to the cpt for X_2, and a_3 is the plausibility of $X_3 = 1$ conditional on $X_1 = 1 \cap X_2 = 0$, given by the cpt for X_3. We do not need to actually define the operation $\otimes$ here; we just leave $a_1 \otimes a_2 \otimes a_3$ as an uninterpreted expression. However, if we have some constraints on the relative order of elements in the cpt (as we do in our examples, and typically will in practice), then lift this to an order on expressions of the form $a_1 \otimes a_2 \otimes a_3$ by taking $a_1 \otimes a_2 \otimes a_3 \leq a'_1 \otimes a'_2 \otimes a'_3$ if and only if, for all a_i, there exists some a'_j such that $a_i \leq a'_j$. The "only if" builds in a minimality assumption: two elements $a_1 \otimes a_2 \otimes a_3$ and $a'_1 \otimes a'_2 \otimes a'_3$ are incomparable unless they are forced to be comparable by the ordering relations among the a_i's and the a'_j's. One advantage of using a partial preorder, rather than a total preorder, is that we can do this.

These assumptions determine a unique partial preorder on elements of the form $a_1 \otimes \cdots \otimes a_n$. This gives us a partial preorder on worlds. (We can then use the construction in Section 3.3 to then obtain an algebraic plausibility measure that in fact is represented by (G, f), but this is no longer necessary, as all we care about is the normality order on worlds.)

While the formal foundations of our approach involve some complexities, the application of these ideas to specific cases is often quite intuitive. The following two examples show how this construction might work in our running example.

Example 3.2

Consider the forest fire example again. Here we can take the worlds to have the form (i, j, k), where i, j, and k are the values of M, L, and FF, respectively. We can represent the independencies in the forest fire example using the network in Fig. 1 (with U removed). Thus, L and M are independent, and FF depends on both of them.

For definiteness, consider the disjunctive case, where either a lightning strike or an arsonist's match suffices for fire. It would be natural to say that lightning strikes and arson attacks are atypical, and that a forest fire typically occurs if either of these events occurs. Suppose that we use d_L^+ to represent the plausibility of $L = 0$ (lightning not occurring) and d_L^- to represent the plausibility of $L = 1$; similarly, we use d_M^+ to represent the plausibility of $M = 0$ and d_M^- to represent the plausibility of $M = 1$. Now the question is what we should take the conditional plausibility of $FF = 0$ and $FF = 1$ to be, given each of the four possible settings of L and M. For simplicity, we take all the four values compatible with the equations to be equally plausible, and have plausibility d_{FF}^+, and the values incompatible with the equations to all be equally plausible and have plausibility d_{FF}^-. This gives us the following cpts:

$$\begin{array}{l}
\mathrm{Pl}(L=0) = d_L^+ > d_L^- = \mathrm{Pl}(L=1) \\
\mathrm{Pl}(M=0) = d_M^+ > d_M^- = \mathrm{Pl}(M=1) \\
\mathrm{Pl}(FF=0 \mid L=0 \wedge M=0) = d_{FF}^+ > d_{FF}^- = \mathrm{Pl}(FF=1 \mid L=0 \wedge M=0) \\
\mathrm{Pl}(FF=1 \mid L=1 \wedge M=0) = d_{FF}^+ > d_{FF}^- = \mathrm{Pl}(FF=0 \mid L=1 \wedge M=0) \\
\mathrm{Pl}(FF=1 \mid L=0 \wedge M=1) = d_{FF}^+ > d_{FF}^- = \mathrm{Pl}(FF=0 \mid L=0 \wedge M=1) \\
\mathrm{Pl}(FF=1 \mid L=1 \wedge M=1) = d_{FF}^+ > d_{FF}^- = \mathrm{Pl}(FF=0 \mid L=1 \wedge M=1).
\end{array} \tag{1}$$

Suppose that we further assume that d_L^+, d_M^+, and d_{FF}^+ are all incomparable, as are d_L^-, d_M^-, and d_{FF}^-. Thus, for example, we cannot compare the degree of typicality of no lightning with that of no arson attacks, or the degree of atypicality of lightning with that of an arson attack. Using the construction above gives us the ordering on worlds given in Fig. 2, where an arrow from w to w' indicates that $w' \succ w$.

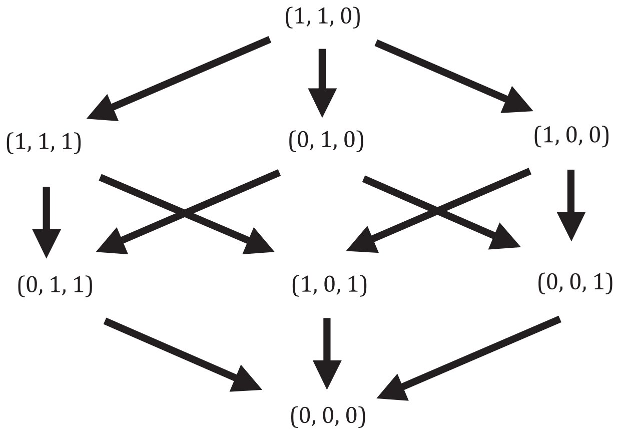

Fig. 2. A normality order on worlds.

In this normality order, (0, 1, 1) is more normal than (1, 1, 1) and (0, 1, 0), but incomparable with (1, 0, 0) and (0, 0, 1). That is because, according to our construction, (0,1,1), (1,1,1), (0,1,0), (1,0,0), and (0,0,1) have plausibility $d_L^+ \otimes d_M^- \otimes d_{FF}^+$, $d_L^- \otimes d_M^- \otimes d_{FF}^+$, $d_L^+ \otimes d_M^- \otimes d_{FF}^+$, $d_L^- \otimes d_M^+ \otimes d_{FF}^-$, and $d_L^+ \otimes d_M^+ \otimes d_{FF}^+$, respectively. The fact that $d_L^+ \otimes d_M^- \otimes d_{FF}^+ \geq d_L^- \otimes d_M^- \otimes d_{FF}^+$ follows as $d_L^+ \geq d_L^-$. The fact that we have >, not just ≥, follows from the fact that we do *not* have $d_L^- \otimes d_M^- \otimes d_{FF}^+ \geq d_L^+ \otimes d_M^- \otimes d_{FF}^+$, as this does not follow from our condition from comparability. The other comparisons follow from similar arguments.

Example 3.3

The order on worlds induced by the Bayesian network in the previous example treats the lightning and the arsonist's actions as incomparable. For example, the world (1, 0, 1), where lightning strikes, the arsonist does not, and there is a fire, is incomparable with the world where lightning does not strike, the arsonist lights his match, and the fire occurs. But this is not the only possibility. Suppose that we judge that it would be more atypical for the arsonist to light a fire than for lightning to strike, and also more typical for the arsonist not to light a fire than for lightning not to strike. (Unlike the case of probability, the latter does not follow from the former.) Recall that this order might reflect the fact that arson is illegal and immoral, rather than the frequency of occurrence of arson as opposed to lightning. While (1) still describes the conditional plausibility tables, we now have $d_L^+ > d_M^+$ and $d_M^- > d_L^-$. This gives us the order on worlds described in Fig. 3.

Now, for example, the world (0,1,1) is strictly more normal than the world (1,0,1); again, the former has plausibility $d_L^+ \otimes d_M^- \otimes d_{FF}^+$, whereas the latter has plausibility $d_L^- \otimes d_M^+ \otimes d_{FF}^+$. But as $d_L^+ > d_L^-$ and $d_L^+ > d_M^+$, by assumption, it follows that $d_L^+ \otimes d_M^- \otimes d_{FF}^+ > d_L^- \otimes d_M^+ \otimes d_{FF}^+$.

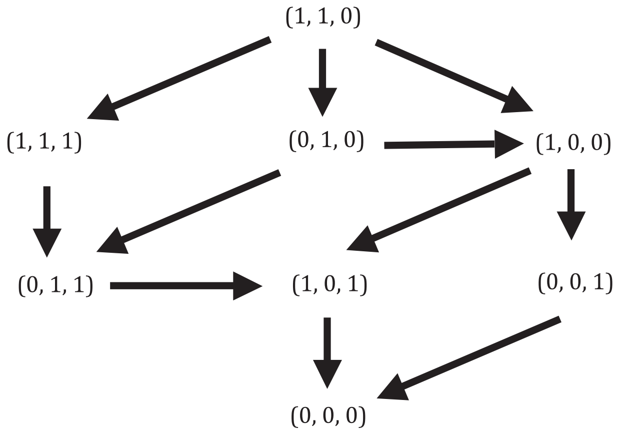

Fig. 3. A different normality order on worlds.

4. Piggy-backing on the causal model

If the normality order is represented by a Bayesian network (G, f), there is no guarantee that the graph G will duplicate the graph corresponding to the causal model. Nonetheless, in many cases it will be reasonable to expect that there will be substantial agreement between the two graphs. When this happens, it will be possible to make parts of the causal model do "double duty": representing both causal structure and the structure of the normality order. In Examples 3.2 and 3.3, the graph describing the causal structure is the same as the graph in the Bayesian network representation of the normality order. This reflects the idea that a fire typically occurs when causes of a fire are present.

But we can say more than this. Consider the conditional plausibility table for the variable FF from (1). We can summarize it as follows:

$$\mathrm{Pl}(FF = ff | L = l \wedge M = m) = \begin{cases} d_{FF^+} & \text{if } ff = \max(l, m) \\ d_{FF^-} & \text{otherwise.} \end{cases}$$

Recall that $FF = \max(L,M)$ is the structural equation for FF in the causal model. So the conditional plausibility table says, in effect, that it is typical for FF to obey the structural equations, and atypical to violate it.

Variables typically obey the structural equations. Thus, it is often far more efficient to assume this holds by default and explicitly enumerate cases where this is not so, rather than writing out all the equations. Specifically, we propose the following default rule.

Default Rule 1 (*Normal Causality*): Let X be a variable in a causal model with no exogenous parents, and let $\mathbf{PA}(X)$ be the vector of parents of X. Let the structural equation for X be $X = f_X(\mathbf{PA}(X))$. Then, unless explicitly given otherwise, there are two plausibility values d_X^+ and d_X^- with $d_X^+ > d_X^-$ such that

$$\mathrm{Pl}(X = x|\mathbf{PA}(X) = \mathbf{px}) = \begin{cases} d_{X^+} & \text{if } x = fx(\mathbf{px}) \\ d_{X^-} & \text{otherwise.} \end{cases}$$

Default Rule 1 tells us that it is typical for variables to satisfy the equations, unless we explicitly stipulate otherwise. In Examples 3.2 and 3.3, *FF* satisfies Default Rule 1. Moreover, it says that, by default, all values of variables that satisfy the equations are equally typical, whereas all those that do not satisfy the equations are equally atypical. Of course, we could allow some deviations from the equations to be more atypical than others; this would be a violation of the default rule. As the name suggests, the default rule is to be assumed, unless explicitly stated otherwise. The hope is that there will be relatively few violations, so there is still substantial representational economy in assuming the rule. That is, the hope is that, once a causal model is given, the normality order can be represented efficiently by providing the conditional plausibility tables for only those variables that violate the default rule, or whose plausibility values are not determined by the default rule (because they have exogenous parents).[9]

The *Normal Causality* rule, by itself, does not tell us how the plausibility values in the cpt for one variable compare with the plausibility values in the cpt for another variable. We therefore supplement our first default rule with a second:

Default Rule 2 (*Minimality*): If d_x and d_y are plausibility values in the conditional plausibility table for distinct variables X and Y and no information is given explicitly regarding the relative orders of d_x and d_y, then d_x and d_y are incomparable. Again, this default rule is assumed to hold only if there is no explicit stipulation to the contrary. Default Rule 2 tells us that the normality ordering among possible worlds should not include any comparisons that do not follow from the equations (via Default Rule 1) together with the information that is explicitly given.[10] In Example 3.2, all three variables satisfy *Minimality*. In Example 3.3, *FF* satisfies *Minimality* with respect to the other two variables, but the variables L and M do not satisfy it with respect to one another (as their values are stipulated to be comparable).

With these two default rules, we can represent the extended causal model in Example 4 succinctly as follows:

$$FF = \max(L, M)$$

$$\mathrm{Pl}(L = 0) > \mathrm{Pl}(L = 1)$$

$$\mathrm{Pl}(M = 0) > \mathrm{Pl}(M = 1).$$

The rest of the structure of the normality order follows from the default rules. In Example 4, we can represent the extended causal model as follows:

$$FF = \max(L,\ M)$$

$$\mathrm{Pl}(M = 0) > \mathrm{Pl}(L = 0) > \mathrm{Pl}(L = 1) > \mathrm{Pl}(M = 1).$$

Again, the rest of the structure follows from the default rules. In each case, the normality order among the eight possible worlds can be represented with the addition of just a few plausibility values to the causal model. Thus, moving from a causal model to an extended causal model need not impose enormous cognitive demands.

Exceptions to the default rules can come in many forms. There could be values of the variables for which violations of the equations are more typical than agreements with the equations. As we suggested, after Default Rule 1, there could be multiple values of typicality, rather than just two for each variable.[11] Or the conditional plausibility values of one variable could be comparable with those of another variable. These default rules are useful to the extent that there are relatively few violations of them. For some settings, other default rules may also be useful; the two rules we have presented are certainly not the only possible useful defaults.

5. Nothing but normality?

In a recent study, Huber (2011) claimed that it is unnecessary to employ distinct modalities for normality and causal structure, and that it is preferable to encompass both in a unified normality structure. Huber's framework employs a family of ranking functions to represent normality. Huber shows that if the ranking functions satisfy a condition that he calls "respect for the equations," one can use the ranking functions as a "similarity metric" on possible worlds and give a semantics for counterfactuals in the spirit of Stalnaker (1968) or Lewis (1973). In this way, all the information about counterfactuals is already contained in the ranking functions; it is unnecessary to give the structural equations as a distinct element of the model. Moreover, this semantics provides truth values for propositions in a richer language than that of Galles and Pearl (1998), Halpern (2000), or Briggs (2012). In particular, it yields truth values for embedded counterfactuals, where the antecedent of a counterfactual conditional includes a counterfactual.

Huber's requirement that the ranking functions respect the equations is similar in spirit to the *Normal Causality* default rule, in that it requires worlds that violate more equations to receive higher rank. (Higher rank corresponds to lower plausibility.) It is a bit more complicated than this, as it also gives priority to worlds where violations occur "later," as measured by number of steps in a directed path. This is supposed to ensure that the closest possible world to w in which some variable X takes a value x different from the one it takes in w is one where X takes the value x due to a "miracle" that occurs as late as possible.

We do not go into all the details of Huber's result here. It is easy to see that if there is a plausibility measure that satisfies *Normal Causality* or something similar, then there is a natural sense in which the structural equations are encoded in that plausibility measure. Huber's result is one specific way of making of this idea precise.

Huber's result is both interesting and technically impressive. Nonetheless, we prefer to retain causal structure and a normality order as distinct modalities. We have this preference for several reasons.

First, while Huber's result provides a kind of conceptual unification, it is not at all clear that it provides a more compact representation. Indeed, as we have argued, it is often possible to provide a representation of the causal structure plus the normality ordering that is very compact.

Second, we think that the normality ordering and the causal model are conceptually representing very different things. The causal structure, as represented in the equations of a causal model, is an objective feature of a system. For example, the accuracy of a causal model can be evaluated by performing appropriate observations and interventions on the system. By contrast, normality can be affected by social rules, moral norms, and the like. The normality order may reflect features of the way in which an agent reasons about a system, but it is not something that can be confirmed experimentally. We believe that actual causation involves both of these components; it is partly objective and partly value-laden. Our framework keeps these two distinct components separate and makes explicit the different roles they play in judgments of actual causation.

Finally, there are examples where normality and causal structure do and should come apart. Huber briefly discusses this point at the very end of his article. He concludes that we should not rely on mere intuitions about normality in cases such as these but should instead put weight on the conceptual economy and unification that result in his framework. As we now show, however, there are some examples where the cleaving of normality and causal structure is justified not only by intuition but also by the demands of a theory of actual causation.

Recall Example 2.1 in which Professor Smith and the administrative assistant took the two remaining pens. We had three endogenous variables: PS, representing whether or not Professor Smith takes a pen; AA, representing whether or not the administrative assistant takes a pen; and PO representing whether or not a problem occurs. To capture the judgments of the subject in the experiment, we want it to be atypical for Professor Smith to take the pen ($PO = 1$). Let us now suppose that we added a further variable to our model: $CP = 1$ if the department chair institutes a policy forbidding faculty members from taking pens; $CP = 0$ if he or she does not institute such a policy. (We could add additional possible values corresponding to alternative policies, such as forbidding everyone from taking pens, but this is not essential to the present point.) How does the new variable CP relate to PS? On the one hand, it seems that CP influences which value of PS is typical. When $CP = 1$, Professor Smith's taking a pen violates a norm. But it is also apparent that the chair's policy had no effect on Professor Smith; he took a pen despite the policy (let us assume that he willfully ignored the policy). Thus, we want our extended model to say *both* that Professor Smith would take the pen if the chair implements the policy, *and* that this violates a norm. We cannot do this if the same ordering is used for both normality and the structural equations.

6. Conclusion

The goals of this study are relatively modest. We highlight a problem that we believe has not been considered in the causality literature and propose a solution to it. We believe that any reasonable approach to causality must pass a minimal "psychological feasibility" test; the models must be representable compactly. We have shown that this can be done in practice with the Halpern–Pearl model and with other approaches that involve structural equations and possibly also a normality ordering. We believe that such compactness considerations should be taken into account in any attempt to model human reasoning; far too often in the philosophical literature, it has not been considered.

Notes

1. Note that U will not typically be a "common cause" in the usual sense. It represents a variety of different factors which need not be correlated. Thus, we do not expect U to induce a correlation between M and L.
2. Again, the fact that X is assigned $U + Y$ (i.e., the value of X is the sum of the values of U and Y) does not imply that Y is assigned $X{-}U$; that is, $F_Y(U,X,Z) = X - U$ does not necessarily hold.
3. This is because the definition of actual causation in Halpern and Pearl (2005) involves worlds generated by performing interventions on a causal model in a fixed *context* (set of values of the exogenous variables). Thus, we need to compare only worlds where the values of exogenous variables are fixed—that is, we are effectively comparing small worlds.
4. Using the full Halpern–Pearl definition of actual causation, a witness world may also incorporate the results of additional interventions besides the intervention on the candidate cause. See Halpern and Hitchcock (2011) for more details.
5. If a potential cause has multiple witnesses, it is graded according to its most normal witness(es).
6. If $\succeq$ was a partial order rather than just a partial preorder, it would satisfy an additional assumption, *antisymmetry*: $s \succeq s'$ and $s' \succeq s$ would have to imply $s = s^{prime}$. This is an assumption we do *not* want to make.
7. Y is a *descendant* of X if there is a directed path from X to Y, where we take X to be a descendant of itself. Y is a nondescendant of X if it is not a descendant of X. We assume that the reader is familiar with the notion of a random variable Y being independent of another random variable X conditional on a set $\mathbf{Z}$ of random variables. See Halpern (2003) and Pearl (1988) for more discussion.
8. Here and subsequently we make use of an artificial terminological convention introduced in Halpern and Hitchcock (2011). We use "typical" and its cognates when talking about individual variables. For example, we say that it is atypical for lightning to strike. We reserve "normal" and its cognates for comparisons of entire worlds. Formally, however, both are represented by plausibility values.

9. It may be possible to formulate more complex versions of Default Rule 1 that accommodate exogenous parents, and allow for more than two default values. We leave these extensions for another occasion.
10. Roughly speaking, in the context of probability, a distribution that maximizes entropy subject to some constraints is one that is (very roughly) the one that makes things "as equal as possible" subject to the constraints. If there are no constraints, it reduces to the classic principle of indifference, which tells us to assign equal probability to different possibilities in the absence of any reason to think some are more probable. In the context of plausibility, where only weak order is assigned, it is possible to push this idea a step further by making the possibilities incomparable.
11. Note that there are many structural equations for a variable X. Indeed, if X has k parents, there are 2^k equations, one for each possible setting of the values of the parents of X. An equation like $FF = \max(L,M)$ packages up the four equations into one compact equation. Default Rule 1 assumes that all agreements with these 2^k equations get a plausibility of d_X^+, and all violations get a plausibility of d_X^-. But we could certainly view some violations as less typical than others.

References

Adams, E. (1975). *The logic of conditionals*. Dordrecht, The Netherlands: Reidel.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies 160*, 139–166.

Dubois, D., & Prade, H. (1991). Possibilistic logic, preferential models, non-monotonicity and related issues. In Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI 91), pp. 419–424.

Friedman, N., & Halpern, J. Y. (1995). Plausibility measures: A user's guide. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 95), pp. 175–184.

Friedman, N., & Halpern, J. Y. (2001). Plausibility measures and default reasoning. *Journal of the ACM 48*(4), 648–685.

Galles, D., & Pearl, J. (1997). Axioms of causal relevance. *Artificial Intelligence 97*(1–2), 9–43.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science 3*(1), 151–182.

Geffner, H. (1992). High probabilities, model preference and default arguments. *Mind and Machines 2*, 51–70.

Goldszmidt, M., & Pearl, J. (1992). Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference (KR 92), pp. 661–672.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies 132*, 109–136.

Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of A.I. Research 12*, 317–337.

Halpern, J. Y. (2001). Conditional plausibility measures and Bayesian networks. *Journal of A.I. Research 14*, 359–389.

Halpern, J. Y. (2003). *Reasoning about uncertainty*. Cambridge, MA: MIT Press.

Halpern, J. Y. (2008). Defaults and normality in causal structures. In Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference (KR 08), pp. 198–208.

Halpern, J. Y., & Hitchcock, C. (2010). Actual causation and the art of modeling. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Causality, probability, and heuristics: A tribute to Judea Pearl*, pp. 383–406. London: College Publications.

Halpern, J. Y., & Hitchcock, C. (2011). Graded causation and defaults. Unpublished manuscript. Available at http://www.cs.cornell.edu/home/halpern/papers/normality.pdf.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science 56*(4), 843–887.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review 116*, 495–532.

Huber, F. (2011). Structural equations and beyond. Unpublished manuscript.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Volume 2: The cognitive science of morality* (pp. 441–447). Cambridge, MA: MIT Press.

Kraus, S. et al. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence 44*, 167–207.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Menzies, P. (2004). Causal models, token causation, and processes. *Philosophy of Science 71*, 820–832.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufmann.

Pearl, J. (1989). *Probabilistic semantics for nonmonotonic reasoning: A survey*. In Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning (KR 89), 505–516. Reprinted in G. Shafer and J. Pearl (Eds.), *Readings in uncertain reasoning*, pp. 699–710. San Francisco: Morgan Kaufmann, 1990.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–710.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. New York: Cambridge University Press.

Shoham, Y. (1987). A semantical approach to nonmonotonic logics. In Proceedings of the 2nd IEEE Symposium on Logic in Computer Science, pp. 275–279. Reprinted in M. L. Ginsberg (Ed.), *Readings in nonmonotonic reasoning*, pp. 227–250. San Francisco: Morgan Kaufman, 1987.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. In W. Harper & B. Skyrms (Eds.), *Causation in decision, belief change, and statistics*, *vol. 2* (pp. 105–134). Dordrecht, The Netherlands: Reidel.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford, England: Blackwell.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, England: Oxford University Press.