Combining Experts' Causal Judgments*

Dalal Alrajeh[a], Hana Chockler[b], Joseph Y. Halpern[c]

[a] *Department of Computing, Imperial College London, UK*
[b] *Department of Informatics, King's College London, UK*
[c] *Computer Science Department, Cornell University, USA*

Abstract

Consider a policymaker who wants to decide which intervention to perform in order to change a currently undesirable situation. The policymaker has at her disposal a team of experts, each with their own understanding of the causal dependencies between different factors contributing to the outcome. The policymaker has varying degrees of confidence in the experts' opinions. She wants to combine their opinions in order to decide on the most effective intervention. We formally define the notion of an effective intervention, and then consider how experts' causal judgments can be combined in order to determine the most effective intervention. We define a notion of two causal models being *compatible*, and show how compatible causal models can be merged. We then use it as the basis for combining experts' causal judgments. We also provide a definition of decomposition for causal models to cater for cases when models are incompatible. We illustrate our approach on a number of real-life examples.

1. Introduction

Consider a policymaker who is trying to decide which intervention, that is, which actions, should be implemented in order to bring about a desired outcome, such as preventing violent behavior in prisons or reducing famine mortality in some country. The policymaker has access to various experts who can advise her on which interventions to consider. Some experts may be (in the policymaker's view) more reliable than others; they may also have different areas of expertise; or may have perceived alternative factors in their analysis. The goal of the policymaker is to choose the best intervention, taking into account the experts' advice.

There has been a great deal of work on combining experts' probabilistic judgments. (**?**] provide a somewhat dated but still useful overview; Dawid [5] and Fenton et al. [7], among others, give a Bayesian analysis.) We are interested in combining experts' judgments in order to decide on the best intervention. Hence, we need more than probabilities. We need to have a causal understanding of the situation. Thus, we assume that the experts provide the policymaker with *causal models*. In general, these models may involve different variables (since the experts may be focusing on different aspects of the

*A preliminary version of the paper appeared in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).*

problem). Even if two models both include variables C_1 and C_2, they may disagree on the relationships between them. For example, one expert may believe that C_2 is independent of C_1 while another may believe that C_1 causally depends on C_2. Yet somehow the policymaker wants to use the information in these causal models to reach her decision.

Despite the clear need for causal reasoning, and the examples in the literature and in practice where experts work with causal models (e.g., [2, 16, 18, 23**?** , 24]), there is surprisingly little work on combining causal judgments. Indeed, the only work that we are aware of that preceded our work is that of Bradley, Dietrich and List [1] (BDL from now on), who prove an impossibility result. Specifically, they describe certain arguably reasonable desiderata, and show that there is no way of merging causal models so as to satisfy all their desiderata. They then discuss various weakenings of their assumptions to see the extent to which the impossibility can be avoided. Our appraoch can be understood as (among other things) weakening two of their assumptions; we discuss this in more detail in Section 3.3. Following the conference version of our paper, Zennaro and Ivanovska [29, 30] examined the problem of merging causal models where the merged model must satisfy a fairness requirement (although the individual experts' models may not be fair). They proposed a way of combining models based on ideas of BDL. Friedenberg and Halpern [8] also considered the same problem of merging causal model of experts, but allowed for the possibility that experts disagree on the causal structure of variables due to having different focus areas. Finally, Feng et al.describe a general method for combining Bayesian networks, that generalizes earlier approach (see [6] and the references therein); their approach is quite different from ours, in part, because they need a way to combine the numerical parameters of the Bayesian networks under consideration.

There is also much work on the closely related problem of *causal discovery*: constructing a single causal model from a data set. A variety of techniques have been used to find the model that best describes how the data is generated (see, e.g., [3, 4, 17, 25, 26]; Triantafillou and Tsamardinos [26] provide a good overview of work in the area).

Of course, if we have the data that the experts used to generate their models, then we should apply the more refined techniques of the work on causal discovery. However, while the causal models constructed by experts are presumably based on data, the data itself is typically no longer available. Rather, the models represent the distillation of years of experience, obtained by querying the experts.

In this paper, we present an approach to merging experts' causal models when sufficient data for discovering the overall causal model is not available. The key step in merging experts' causal models lies in defining when two causal models are *compatible*. Causal models can be merged only if they are compatible. We start with a notion of *strong* compatibility, where the conditions say, among other things, that if both M_1 and M_2 involve variables C_1 and C_2, then they must agree on the causal relationship between C_1 and C_2. But that is not enough. Suppose that in both models C_1 depends on C_2, C_3, and C_4. Then in a precise sense, the two models must agree on *how* the dependence works, despite describing the world using possibly different sets of variables. Roughly speaking, this is the case when, for every variable C that the two models have in common, we can designate one of the models as being "dominant" with respect to C, and use that model to determine the relationships for C. When M_1 and M_2 are compatible, we are able to construct a merged model $M_1 \oplus M_2$ that can be viewed as satisfying all but one of BDL's desiderata (and we argue that the one it does not satisfy is unreasonable).

In a precise sense, all conclusions that hold in either of the models M_1 and M_2 also

hold in the merged model (see Theorem 3.13(e)). In this way, the merged model takes advantage of the information supplied by all the experts (at least, to the extent that the experts' models are compatible), and can go beyond what we can do with either of the individual models (e.g., considering interventions that simultaneously act on variables that are in M_1 but not in M_2 and variables that are in M_2 but not in M_1).

The set of constraints that need to be satisfied for models to be compatible may be restrictive in some cases; as we show on real-life examples, models are often not compatible, due to disagreements about some parts of the model, even though some interventions being considered do not affect those parts of the model. We therefore introduce a notion of causal model decomposition to allow policymakers to "localize" the incompatibility between models, and merge the parts of the models that are compatible.

Having set out the formal foundation for merging causal models, we show how probabilities can be assigned to different reasonable ways of merging experts' causal models based on the perceived reliability of the experts who proposed them, using relatively standard techniques. The policymaker will then have a probability on causal models that she can use to decide which interventions to implement. Specifically, we can use the probability on causal models to compute the probability that an intervention is efficacious. Combining that with the cost of implementing the intervention, the policymaker can compute the most effective intervention. As we shall see, although we work with the same causal structures used to define causality, interventions are different from (and actually simpler to analyze than) causes.

We draw on various examples from the literature (including real-world scenarios involving complex sociological phenomena) to illustrate our approach, including crime-prevention scenarios [**?**], radicalization in prisons [**?**], and child abuse [**?**]. These examples reinforce our belief that our approach provides a useful formal framework that can be applied to the determination of appropriate interventions for policymaking.

The rest of the paper is organized as follows. Section 2 provides some background material on causal models. We formally define our notion of intervention and compare it to causality in Section 3. We discuss our concept of compatibility and how causal models can be decomposed and merged in Section **??**. We discuss how the notions of interventions and of compatible models can be used by the policymakers to choose optimal interventions in Section 4. Finally, we summarize our results and outline future directions in Section 5.

2. Causal Models

In this section, we review the definition of causal models introduced by Halpern and Pearl [13]. The material in this section is largely taken from [11].

We assume that the world is described in terms of variables and their values. Some variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the 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. For example, in a voting scenario, we could have endogenous variables that describe what the voters actually do (i.e., which candidate they vote for), exogenous variables that describe the factors that determine how the voters vote, and a variable describing the outcome (who wins). The structural equations describe how these values are determined

(majority rules; a candidate wins if A and at least two of B, C, D, and E vote for him; etc.).

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 $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (that is, the set of values over which Y *ranges*). For simplicity, we assume here that $\mathcal{V}$ is finite, as is $\mathcal{R}(Y)$ for every endogenous variable $Y \in \mathcal{V}$. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X (i.e., $F_X = \mathcal{F}(X)$) 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 $X = U + Y$. Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.[1]

The structural equations define what happens in the presence of external interventions. Setting the value of some variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{X \leftarrow x}$, which is identical to M, except that the equation F_X for X in $\mathcal{F}$ is replaced by $X = x$.

The dependencies between variables in a causal model M can be described using a *causal network* (or *causal graph*), whose nodes are labeled by the endogenous and exogenous variables in $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$, with one node for each variable in $\mathcal{U} \cup \mathcal{V}$. The roots of the graph are (labeled by) the exogenous variables and endogenous variables X such that F_X to Y if Y *depends on* X; this is the case if there is some setting of all the variables in $\mathcal{U} \cup \mathcal{V}$ other than X and Y such that varying the value of X in that setting results in a variation in the value of Y; that is, there is a setting $\vec{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x, \vec{z}) \neq F_Y(x', \vec{z})$. A causal model M is *recursive* (or *acyclic*) if its causal graph is acyclic. It should be clear that if M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, the values of all the other variables are determined (i.e., there is a unique solution to all the equations). We can determine these values by starting at the top of the graph and working our way down. What we are calling here "recursive" is called *strongly recursive* by Halpern [11], who reserves the term "recursive" for a model where, for each context $\vec{u}$, the dependency graph is acyclic (but it may be a different acyclic graph for context, so that in one context A may be an ancestor of B, while in another, B may be an ancestor of A). In this paper, following most of the rest of the literature (see, e.g. [11]), we restrict for simplicity to strongly recursive models, although our main definitions and apply with only minor changes to Halpern's context-dependent notion of recursivity. (We explain the changes needed at the relevant points.)

The following example, due to Lewis [21], describes a simple causal scenario.

[1]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 that is, $F_Y(U, X, Z) = X - U$ does not necessarily hold. The assignment describes the effect of interventions. While intervening on Y or U might affect X, intervening on X might not affect Y. Indeed, if X causally depends on U and Y, then Y does not in general depend on X.

Example 2.1. Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw. A naive model might have an exogenous variable U that encapsulates whatever background factors cause Suzy and Billy to decide to throw the rock (the details of U do not matter, since we are interested only in the context where U's value is such that both Suzy and Billy throw), a variable ST for Suzy throws ($ST = 1$ if Suzy throws, and $ST = 0$ if she doesn't), a variable BT for Billy throws, and a variable BS for bottle shatters. In the naive model, whose graph is given in Figure 1, BS is 1 if one of ST and BT is 1. Thus, U has four possible values, depending on which of Suzy and Billy throw. We also have three binary variables: ST for Suzy throws, BT for Billy throws, and BS for bottle shatters. $ST = 1$ means "Suzy throws"; $ST = 0$ means that she does not. We interpret $BT = 1$, $BT = 0$, $BS = 1$, and $BS = 0$ similarly. The values of ST and BT are determined by the context. The value of BS is determined by the equation $F_{BS}(\vec{u}, ST, BT) = ST \vee BT$. The causal graph corresponding to this model is depicted in Figure 1.

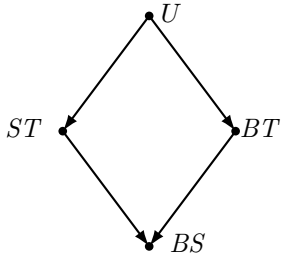


Figure 1: A naive model for the rock-throwing example.

This causal model does not distinguish between Suzy and Billy's rocks hitting the bottle simultaneously and Suzy's rock hitting first. A more sophisticated model might also include variables SH and BH, for Suzy's rock hits the bottle and Billy's rock hits the bottle. It is immediate from the equations that BS is 1 iff one of SH and BH is 1. However, now, SH is 1 if ST is 1, and $BH = 1$ if $BT = 1$ and $SH = 0$. Thus, Billy's throw hits if Billy throws *and* Suzy's rock doesn't hit. This model is described by the graph in Figure 2, where we implicitly assume a context where Suzy throws first, so there is an edge from SH to BH, but not one in the other direction (and omit the exogenous variable). [2] ∎

In several papers in the literature (e.g., [1**?**]), a causal model is defined simply by a causal graph indicating the dependencies, perhaps also showing whether a change has a positive or negative effect; that is, edges are annotated with + or −, so that an edge from A to B annotated with + means that an increase in A results in an increase in B, while if it is annotated with a −, then an increase in A results in a decrease in B (where what constitutes an increase or decrease is determined by the model). Examples of these

[2]We remark that if we allowed who hits first to depend on the context, we would get a context-dependent recursive model in the sense of Halpern [11], where the direction of the arrow from SH to BH would depend on the context.

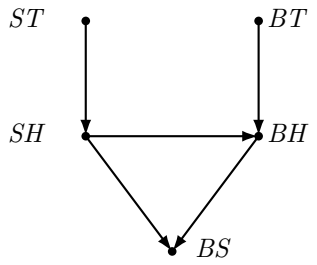


Figure 2: A better model for the rock-throwing example.

are shown in Section **??**. Our models are more expressive, since the equations typically provide much more detailed information regarding the dependence between variables (as shown in Example 2.1); the causal graphs capture only part of this information. Of course, this extra information makes merging models more difficult (although, as the results of BDL show, the difficulties in merging models already arise with purely qualitative graphs).

To define interventions carefully, it is useful to have a language in which we can make statements about interventions. Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where φ is a Boolean combination of primitive events, $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{U} \cup \mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$.[3] Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

We call a pair $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$ a *(causal) setting*. A causal formula ψ is true or false in a setting. We write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \vec{u})$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$. (i.e., the unique vector of values for the exogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). If $k \geq 1$ and Y_k is an endogenous variable, then

$$(M, \vec{u}) \models [Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi \text{ iff}$$
$$(M_{Y_k \leftarrow y_k}, \vec{u}) \models [Y_1 \leftarrow y_1, \ldots, Y_{k-1} \leftarrow y_{k-1}]\varphi.$$

If Y_k is an exogenous variable, then

$$(M, \vec{u}) \models [Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi \text{ iff}$$
$$(M, \vec{u}[Y_k/y_k]) \models [Y_1 \leftarrow y_1, \ldots, Y_{k-1} \leftarrow y_{k-1}]\varphi,$$

where $\vec{u}[Y_k/y_k]$ is the result of replacing the value of Y_k in $\vec{u}$ by y_k.

[3] In earlier work [11, 13], each Y_i was taken to be an endogenous variable. For technical reasons (explained in Section **??**), we also allow Y to be exogenous.

3. Interventions

In this section, we define (causal) interventions, and compare the notion of intervention to that of cause.

Definition 3.1. *[Intervention leading to $\neg\varphi$]* $\vec{X} = \vec{x}$ *is an* intervention leading to $\neg\varphi$ in $(M, \vec{u})$ *if the following three conditions hold:*

I1. $(M, \vec{u}) \models \varphi$.

I2. $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\neg\varphi$.

I3. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *and values* $\vec{x}'$ *such that* $\vec{X}' = \vec{x}'$ *satisfies I2.*

I1 says φ must be true in the current setting $(M, \vec{u})$, while I2 says that performing the intervention results in φ no longer being true. I3 is a minimality condition. From a policymaker's perspective, I2 is the key condition. It says that by making the appropriate changes, we can bring about a change in φ.

Our definition of intervention is essentially equivalent to others in the literature. Pearl [22**?**] assumes that the causal model is first analyzed, and then a new intervention variable I_V is added for each variable V on which we want to intervene. If $I_V = 1$, then the appropriate intervention on V takes place, independent of the values of the other parents of V; if $I_V = 0$, then I_V has no effect, and the behavior of V is determined by its parents, just as it was in the original model. **?**], **?**], and Woodward [28] take similar approaches. If $\vec{X}$ consists of the variables $\{X_1, \ldots, X_k\}$, then to model the intervention $\vec{X} \leftarrow \vec{x}$ in this framework, we would also have to set the variables $I_{X_1}, \ldots, I_{X_k}$ to 1.

We do not require special intervention variables; we just allow interventions directly on the variables in the model. But we can certainly assume as a special case that for each variable V in the model there is a special intervention variable I_V that works just like Pearl's intervention variables, and thus recover the other approaches considered in the literature. All these definitions are trying to capture similar intuitions, and each approach can capture the others. Definition 3.1 focuses on the outcome of the intervention, not just the intervention itself, since this is what we will be most interested in in this paper.

Although there seems to be relatively little disagreement about how to capture intervention, the same cannot be said for causality. Even among definitions that involve counterfactuals and structural equations [9, 10, 13, 14, 15, 28], there are a number of subtle variations. Fortunately for us, the definition of intervention does not depend on how causality is defined. While we do not get into the details of causality here, it is instructive to compare the definitions of causality and intervention.

For definiteness, we focus on the definition of *actual causality* given by Halpern [10]. It has conditions AC1–3 that are analogues of I1–3. Specifically, AC1 says $\vec{X} = \vec{x}$ is a cause of φ in $(M, \vec{u})$ if $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$ and AC3 is a minimality condition. AC2 is a more complicated condition; it says that there exist values $\vec{x}'$ for the variables in $\vec{X}$, a (possibly empty) subset $\vec{W}$ of variables, and values $\vec{w}$ for the variables in $\vec{W}$ such that $(M, \vec{u}) \models \vec{W} = \vec{w}$ and $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}]\neg\varphi$. We do not attempt to explain or motivate AC2 here, since our focus is not causality.

Consider Example 2.1 again. Changing the value of Suzy's throw by itself is not an intervention leading to the bottle not being shattered. Even if we prevent Suzy from

throwing, the bottle will still shatter because of Billy's throw. That is, although $ST = 1$ is a cause of the bottle shattering, $ST = 0$ is not an intervention leading to the bottle not being shattered; intervening on ST alone does not change the outcome. On the other hand, $ST = 0 \wedge BT = 0$ is an intervention leading to the bottle not being shattered, but $ST = 1 \wedge BT = 1$ is not a cause of the bottle shattering; it violates the minimality condition AC3.

It is almost immediate from the definitions that we have the following relationship between interventions and causes:

Proposition 3.2. *If $\vec{X} = \vec{x}$ is an intervention leading to $\neg\varphi$ in $(M, \vec{u})$, then there is some subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ is a cause of φ in $(M, \vec{u})$, where $\vec{x}'$ is such that $(M, \vec{u}) \models \vec{X}' = \vec{x}'$. Conversely, if $\vec{X} = \vec{x}$ is a cause of φ in $(M, \vec{u})$ then there is a superset $\vec{X}'$ of $\vec{X}$ and values $\vec{x}'$ such that $\vec{X}' = \vec{x}'$ is an intervention leading to $\neg\varphi$.*

Halpern [10] proved that (for his latest definition) the complexity of determining whether $\vec{X} = \vec{x}$ is a cause of φ in $(M, \vec{u})$ is DP-complete, where DP consists of those languages L for which there exist a language L_1 in NP and a language L_2 in co-NP such that $L = L_1 \cap L_2$ [**?**]. It is well known that DP is at least as hard as NP and co-NP (and most likely strictly harder). The following theorem shows that the problem of determining whether $\vec{X} = \vec{x}$ is an intervention is in a lower complexity class.

Theorem 3.3. *Given a causal model M, a context $\vec{u}$, and a Boolean formula φ, the problem of determining whether $\vec{X} = \vec{x}$ is an intervention eading to $\neg\varphi$ in $(M, \vec{u})$ is co-NP-complete.*

Proof. First, we prove membership in co-NP. It is easy to see that checking conditions I1 and I2 of Definition 3.1 can be done in polynomial time by simply evaluating φ first in $(M, \vec{u})$ and then in the modified context where the values of $\vec{X}$ are set to $\vec{x}$. Checking whether I3 holds is in co-NP, because the complementary condition is in NP; indeed, we simply have to guess a subset $\vec{X}'$ of $\vec{X}$ and values $\vec{x}'$ and verify that I1 and I2 hold for $\vec{X}' = \vec{x}'$ and φ, which, as we observed, can be done in polynomial time.

For co-NP-hardness, we provide a reduction from UNSAT, which is the language of all unsatisfiable Boolean formulas, to the intervention problem. Given a formula ψ that mentions the set $\vec{V}$ of variables, we construct a causal model M_ψ, context $\vec{u}$, and formula φ such that $\vec{V} = 1$ is an intervention leading to $\neg\varphi$ in $(M, \vec{u})$ iff ψ is unsatisfiable.

The set of endogenous variables in M is $\vec{V} \cup \{V', Y\}$, where V' and Y are fresh variables not in $\vec{V}$. Let $\vec{W} = \vec{V} \cup \{V'\}$. There is a single exogenous variable U that determines the value of the variables in $\vec{W}$: we have the equation $V = U$ for each variable $V \in \vec{W}$. The equation for Y is $Y = \vee_{V \in \vec{W}} (V = 0)$ (so $Y = 1$ if at least one variable in $\vec{W}$ is 0). Let φ be $\neg\psi \wedge (Y = 1)$. We claim that $\vec{W} = \vec{1}$ is an intervention leading to $\neg\varphi$ in $(M_\psi, 0)$ iff $\psi \in$ UNSAT.

Suppose that $\psi \in$ UNSAT. Then, it is easy to see that $(M, 0) \models \varphi$ (since $\neg\psi$ is valid) and $(M, 0) \models [\vec{W} \leftarrow \vec{1}]\neg\varphi$ (since $(M, 0) \models [\vec{W} \leftarrow \vec{1}](Y = 0)$). To see that I3 holds, suppose by way of contradiction that $\vec{W}' \leftarrow \vec{w}'$ satisfies I1 and I2 for some strict subset $\vec{W}'$ of $\vec{W}$. In particular, we must have $(M, 0) \models [\vec{W}' \leftarrow \vec{w}']\neg\varphi$. We clearly have $(M, 0) \models [\vec{W}' \leftarrow \vec{w}'](Y = 1)$, so we must have $(M, 0) \models [\vec{W}' \leftarrow \vec{w}']\psi$, contradicting the assumption that $\psi \in$ UNSAT. Thus, $\vec{W} \leftarrow \vec{1}$ is an ntervention leading to $\neg\varphi$, as desired.

For the converse, suppose that $\vec{W} \leftarrow \vec{1}$ is an intervention leading to $\neg\varphi$. Then we must have $(M, 0) \models [\vec{W}' \leftarrow \vec{w}']\neg\psi$ for all strict subsets $\vec{W}'$ of $\vec{W}$ and all settings $\vec{w}'$ of the variables in $\vec{W}'$. Since, in particular, this is true for all subsets $\vec{W}'$ of $\vec{W}$ that do not involve V', it follows that $\neg\psi$ is true for all truth assignments, so $\psi \in$ UNSAT. ∎

In practice, however, we rarely expect to face the co-NP complexity. For reasons of cost or practicality, we would expect a policymaker to consider interventions on at most k variables, for some small k. The straightforward algorithm that, for a given k, checks all sets of variables of the model M of size at most k runs in time $O(|M|^k)$.

This section presents our definition for compatibility of expert opinions. We consider each expert's opinion to be represented by a causal model and, for simplicity, that each expert expresses her opinion with certainty. (We can easily extend our approach to allow the experts to have some uncertainty about the correct model; see the end of Section 4.) We then formalize the notion of decomposition of causal models, and show how this enables portions of incompatible models to be combined.

3.1. Compatibility

To specify what it means for a set of models to be compatible, we first define what it means for the causal model M_1 to contain at least as much information about variable C as the causal model M_2, denoted $M_1 \succeq_C M_2$. Intuitively, M_1 contains at least as much information about C as M_2 if M_1 and M_2 say the same things about the causal structure of C, but M_1 contains (possibly) more information about C, because, for example, there are additional variables in M_1 that affect C. We capture this property formally below. We say that B is an *immediate M_2-ancestor of Y in M_1* if $B \in \mathcal{U}_2 \cup \mathcal{V}_2$, B is an ancestor of Y in M_1, and there is a path from B to Y in M_1 that has no nodes in $\mathcal{U}_2 \cup \mathcal{V}_2$ other than B and Y (if $Y \in \mathcal{U}_2 \cup \mathcal{V}_2$). That is, Y is the first node in M_2 after B on a path from B to Y in M_1.

Definition 3.4. *[Strong domination] Let $M_1 = ((\mathcal{U}_1, \mathcal{V}_1, \mathcal{R}_1), \mathcal{F}_1)$ and $M_2 = ((\mathcal{U}_2, \mathcal{V}_2, \mathcal{R}_2), \mathcal{F}_2)$. Let $Par_M(C)$ denote the variables that are parents of C in M. M_1* strongly dominates *M_2 with respect to C, denoted $M_1 \succeq_C M_2$, if the following conditions hold:*

MI1$_{M_1,M_2,C}$**.** *The parents of C in M_2, $Par_{M_2}(C)$, are the immediate M_2-ancestors of C in M_1.*

MI2$_{M_1,M_2,C}$**.** *Every path from an exogenous variable to C in M_1 goes through a variable in $Par_{M_2}(C)$.*

MI3$_{M_1,M_2,C}$**.** *Let $X = ((\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2)) - \{C\}$. Then for all settings $\vec{x}$ of the variables in $\vec{X}$, all values c of C, all contexts $\vec{u}_1$ for M_1, and all contexts $\vec{u}_2$ for M_2, we have*

$$(M_1, \vec{u}_1) \models [\vec{X} \leftarrow \vec{x}](C = c) \text{ iff } (M_2, \vec{u}_2) \models [\vec{X} \leftarrow \vec{x}](C = c).$$

If MI1$_{M_1,M_2,C}$ holds and, for example, B is a parent of C in M_2, then there may be a variable B' on the path from B to C in M_1. Thus, M_1 has in a sense more detailed information than M_2 about the causal paths leading to C. MI1$_{M_1,M_2,C}$ is not by itself enough to say that M_1 and M_2 agree on the causal relations for C. This is guaranteed by

$\text{MI2}_{M_1,M_2,C}$ and $\text{MI3}_{M_1,M_2,C}$. $\text{MI2}_{M_1,M_2,C}$ says that the variables in $Par_{M_2}(C)$ screen off C from the exogenous variables in M_1. (Clearly the variables in $Par_{M_2}(C)$ also screen off C from the exogenous variables in M_2.) It follows that if $(M_1, \vec{u}_1) \models [Par_{M_2}(C) \leftarrow \vec{x}](C = c)$ for some context $\vec{u}_1$, then $(M_1, \vec{u}) \models [Par_{M_2}(C) \leftarrow \vec{x}](C = c)$ for all contexts $\vec{u}$ in M_1, and similarly for M_2. In light of this observation, it follows that $\text{MI3}_{M_1,M_2,C}$ assures us that C satisfies the same causal relations in both models. We write $M_1 \not\succeq_C M_2$ if any of the conditions above does not hold.

Two technical comments regarding Definition 3.4: First, note that in MI3 we used the fact that we allow the $\vec{X}$ in formulas of the form $[\vec{X} \leftarrow \vec{x}]\varphi$ to include exogenous variables, since some of the parents of C may be exogenous. Second, despite the suggestive notation, $\succeq_C$ is not a partial order. In particular, it is not hard to construct examples showing that it is not transitive. However, $\succeq_C$ is a partial order on compatible models (see the proof of Proposition 3.13), which is the only context in which we are interested in transitivity, so the abuse of notation is somewhat justified.

Note that we have a relation $\succeq_C$ for each variable C that appears in both M_1 and M_2. Model M_1 may be more informative than M_2 with respect to C whereas M_2 may be more informative than M_1 with respect to another variable C'. Roughly speaking, M_1 and M_2 are *compatible* if for each variable $C \in \mathcal{V}_1 \cap \mathcal{V}_2$, either $M_1 \succeq_C M_2$ or $M_2 \succeq_C M_1$. We then merge M_1 and M_2 by taking the equations for C to be determined by the model that has more information about C. Consider another example demonstrating the notion of strong dominance,

Example 3.5. [1] An aid agency consults two experts about causes of famine in a region. Both experts agree that the amount of rainfall (R) affects crop yield (Y). Specifically, a shortage of rainfall leads to poor crop yield. Expert 2 says that political conflict (P) can also directly affect famine. Expert 1, on the other hand, says that P affects F only via Y. The experts' causal graphs are depicted in Figure 3, where the graph on the left, M_1, describes expert 1's model, while the graph on the right, M_2, describes expert 2's model. These graphs already appear in the work of BDL. We show only the structural equations where the two experts differ in their opinions. In these graphs (as in many other causal graphs in the literature), the exogenous variables are typically omitted; unless we explicitly say otherwise, all the variables are taken to be endogenous. Neither $\text{MI1}_{M_1,M_2,F}$

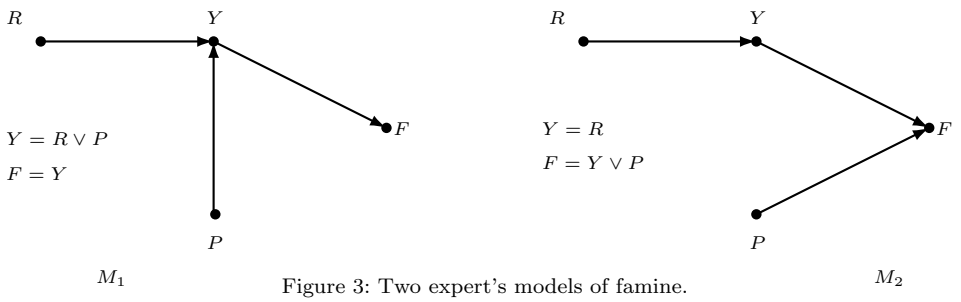


Figure 3: Two expert's models of famine.

nor $\text{MI1}_{M_2,M_1,F}$ holds, since P is not an M_2-immediate ancestor of F in M_1. Similarly, neither $\text{MI1}_{M_1,M_2,Y}$ nor $\text{MI1}_{M_2,M_1,Y}$ holds, since P is not an M_1-immediate ancestor of Y in M_2 (indeed, it is not an ancestor at all). $\text{MI2}_{M_1,M_2,F}$ holds since every path in M_1

from an exogenous variable to F goes through a variable that is a parent of F in M_2 (namely, Y); $\text{MI2}_{M_2,M_1,F}$ does not hold (there is a path in M_2 to F via P that does not go through a parent of F in M_1). Although we are not given the equations, we also know that $\text{MI3}_{M_1,M_2,F}$ does not hold. Since P is a parent of F in M_2 according to expert 2, there must be a setting y of Y such that the value of F changes depending on the value of P if $Y = y$. This cannot be the case in M_1, since Y screens off P from F. It easily follows that taking $\vec{X} = (P, Y)$ we get a counterexample to $\text{MI3}_{M_1,M_2,F}$. Therefore, we have neither $M_1 \succeq_F M_2$ nor $M_2 \succeq_F M_1$. ∎

While the definition of dominance given above is useful, it does not cover all cases where a policymaker may want to merge models. Consider the following example, taken from the work of **?**].

Example 3.6. Two experts have provided causal models regarding the causes of domestic violence. According to the first expert, an appropriate arrest policy (*AP*) may affect both an offender's belief that his partner would report any abuse to police (*PLS*) and the amount of domestic violence (*DV*). The amount of domestic violence also affects the likelihood of a victim calling to report abuse (*C*), which in turn affects the likelihood of there being a random arrest (*A*). (Decisions on whether to arrest the offender in cases of domestic violence were randomized.)

According to the second expert, *DV* affects *A* directly, while *A* affects the amount of repeated violence (*RV*) through both formal sanction (*FS*) and informal sanction on socially embedded individuals (*IS*). **?**] use the causal graphs shown in Figure 4, which are annotated with the direction of the influence (the only information provided by the experts) to describe the expert's opinions.

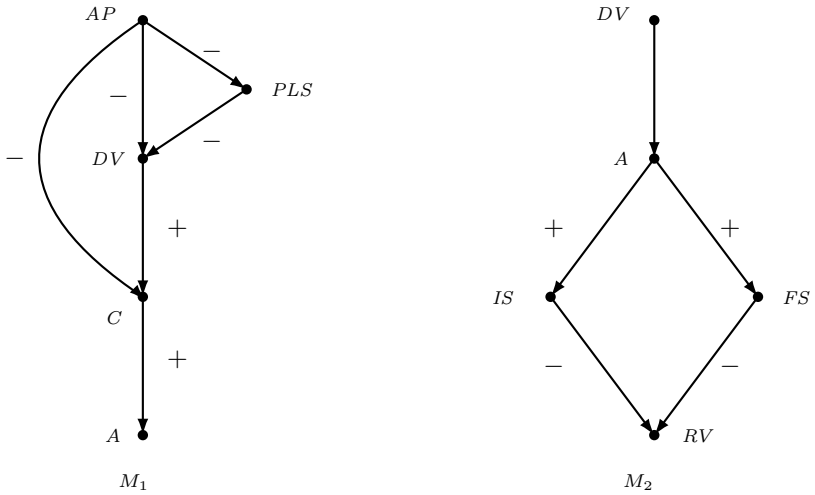


Figure 4: Experts' models of domestic violence.

For the two common variables *DV* and *A*, $\text{MI1}_{M_1,M_2,DV}$ and $\text{MI1}_{M_1,M_2,A}$ both hold. If the only variables that have exogenous parents are *AP* in M_1 and *DV* in M_2, and the set of parents of *AP* in M_1 is a subset of the set of parents of *DV* in M_2, then $\text{MI2}_{M_1,M_2,DV}$

holds. Sampson et al. seem to be implicitly assuming this, and that MI3 holds, so they merge M_1 and M_2 to get the causal graph shown in Figure 5.

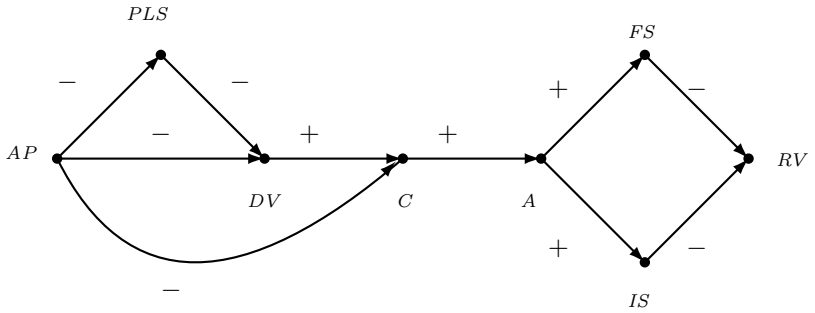


Figure 5: The result of merging experts' model of domestic violence.

Sampson et al. do not provide structural equations. Moreover, for edges that do not have a + or − annotation, such as the edge from DV to A in Figure 4, we do not even know qualitatively what the impact of interventions is. Presumably, the lack of annotation represents the expert's uncertainty. We can capture this uncertainty by viewing the expert as having a probability on two models that disagree on the direction of DV's influence on A (and thus are incompatible because they disagree on the equations). We discuss in Section 4 how such uncertainty can be handled. ∎

To get a more general notion of domination, it turns out to be useful to work at the level of causal settings rather than causal models.

Definition 3.7. *[Weak domination]* $(M_1, \vec{u}_1)$ weakly dominates $(M_2, \vec{u}_2)$ *with respect to* C*, denoted* $(M_1, \vec{u}_1) \succeq_C (M_2, \vec{u}_2)$*, if* $MI1_{M_1,M_2,C}$ *holds,*[4] *and, in addition, the following condition (which can be viewed as a replacement for* $MI2_{M_1,M_2,C}$ *and* $MI3_{M_1,M_2,C}$*) holds:*

MI4$_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ *Let* $\vec{X} = (\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2) - \{C\}$*. Then for all settings* $\vec{x}$ *of the variables in* $\vec{X}$ *and all values* c *of* C*, we have*

$$(M_1, \vec{u}_1) \models [\vec{X} \leftarrow \vec{x}](C = c) \text{ iff } (M_2, \vec{u}_2) \models [\vec{X} \leftarrow \vec{x}](C = c).$$

Lemma 3.8. *If* $M_1 \succeq_C M_2$ *(as defined in Definition 3.4), then for all settings* $(M_1, \vec{u}_1)$ *and* $(M_2, \vec{u}_2)$*, we have that* $(M_1, \vec{u}_1) \succeq_C (M_2, \vec{u}_2)$*.*

Proof. Suppose that $M_1 \succeq_C M_2$. Clearly $\text{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ is a special case of $\text{MI3}_{M_1,M_2,C}$. Thus, $(M_1, \vec{u}_1) \succeq_C (\mathcal{M}_2, \vec{u}_2)$. ∎

In light of Lemma 3.8, we give all the definitions in the remainder of the paper using weak domination. All the technical results hold if we replace weak domination by strong domination.

One more observation will be useful to motivate our definition:

[4] If we consider models that are recursive in the sense of Halpern [11], then we must replace $\text{MI1}_{M_1,M_2,C}$ by $\text{MI1}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$, which is identical except that we need to consider the parents of C in $(M_2, \vec{u}_2)$ rather than M_2 (since now the dependence relation can depend on the context), and the immediate M_2-ancestors of C in $(M_1, \vec{u}_1)$. With this change, the definition applies to models that satisfy context-dependent recursivity.

Lemma 3.9. *If $C \in V_1 \cap V_2$, then there exists contexts $\vec{u}_1$, $\vec{u}'_1$, $\vec{u}_2$, and $\vec{u}'_2$ such that $(M_1, \vec{u}_1) \succeq_C (M_2, \vec{u}_2)$ and $(M_2, \vec{u}'_2) \succeq_C (M_1, \vec{u}'_1)$, then $(M_1, \vec{u}'_1) \succeq_C (M_1, \vec{u}'_2)$.*

Proof. Since $(M_1, \vec{u}_1) \succeq_C (M_2, \vec{u}_2)$, $\mathrm{MI1}_{M_1,M_2,C}$ holds. Since $(M_2, \vec{u}'_2) \succeq_C (M_1, \vec{u}'_1)$, $\mathrm{MI4}_{(M_2,\vec{u}'_2),(M_1,\vec{u}'_1),C}$ holds. Clearly, $\mathrm{MI4}_{(M_1,\vec{u}'_1),(M_2,\vec{u}'_2),C}$ also holds. Hence, $(M_1, \vec{u}'_1) \succeq_C (M_2, \vec{u}'_2)$. ∎

We remark that this lemma does not hold for Halpern's context-dependent notion of recursivity.

Suppose that we have two models M_1 and M_2 such that X is exogenous in M_2 and endogenous in M_1 (as is the case, for example, for the variables A and DV in Figure 4). We might hope that M_1 somehow dominates M_2 with respect to X; intuitively, M_1 has more information about X because it can explain the value of X as due to the values of other variables. However, we cannot hope to show that $(M_1, \vec{u}_1) \succeq_X (M_2, \vec{u}_2)$ for all contexts $\vec{u}_1$ and $\vec{u}_2$. For suppose that $(M_1, \vec{u}_1) \models X = x$. Then unless $X = x$ in $\vec{u}_2$, we do not have $(M_2, \vec{u}_2) \models X = x$. It easily follows that $\mathrm{MI4}_{M_1,\vec{u}_1,M_2,\vec{u}_2,C}$ does not hold. This motivates the following definition.

Definition 3.10. *[Compatibility of causal settings] Causal settings $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are* compatible *if (1) $\vec{u}_1$ and $\vec{u}_2$ agree on all variables in $\mathcal{U}_1 \cap \mathcal{U}_2$, (2) for all variables $X \in \mathcal{U}_1 \cap \mathcal{V}_2$, we have that $X = x$ in $\vec{u}_1$ iff $(M_2, \vec{u}_2) \models X = x$ and (3) for all $X \in \mathcal{U}_2 \cap \mathcal{V}_1$, we have that $X = x$ in $\vec{u}_2$ iff $(M_1, \vec{u}_1) \models X = x$.*

The following definition formalizes compatibility of causal models, independent of the context.

Definition 3.11. *[Compatibility of causal models] If $M_1 = ((\mathcal{U}_1, \mathcal{V}_1, \mathcal{R}_1), \mathcal{F}_1)$ and $M_2 = ((\mathcal{U}_2, \mathcal{V}_2, \mathcal{R}_2), \mathcal{F}_2)$, then M_1 and M_2 are* compatible *if (1) there exist $\vec{u}_1$ and $\vec{u}_2$ such that the causal settings $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are compatible; (2) for all variables $C \in (\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2)$, we have $\mathcal{R}_1(C) = \mathcal{R}_2(C)$; and (3) for all variables $C \in (\mathcal{V}_1 \cap \mathcal{V}_2) \cup (\mathcal{V}_1 \cap \mathcal{U}_2) \cup (\mathcal{V}_2 \cap \mathcal{U}_1)$, we have that either $(M_1, \vec{u}_1) \succeq_C (M_2, \vec{u}_2)$ for all compatible causal settings $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ or $(M_2, \vec{u}_2) \succeq_C (M_1, \vec{u}_1)$ for all compatible causal settings $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$.*

We now can define the result of merging two compatible models.

Definition 3.12. *[Merging compatible models] If $M_1 = ((\mathcal{U}_1, \mathcal{V}_1, \mathcal{R}_1), \mathcal{F}_1)$ and $M_2 = ((\mathcal{U}_2, \mathcal{V}_2, \mathcal{R}_2), \mathcal{F}_2)$ are compatible, then the* merged model $M_1 \oplus M_2$ *is the causal model $((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$, where*

- $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2 - (\mathcal{V}_1 \cup \mathcal{V}_2)$;
- $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$;
- *if $C \in \mathcal{U}_1 \cup \mathcal{V}_1$, then $\mathcal{R}(C) = \mathcal{R}_1(C)$, and if $C \in \mathcal{U}_2 \cup \mathcal{V}_2$, then $\mathcal{R}(C) = \mathcal{R}_2(C)$;*
- *if $C \in \mathcal{V}_1 - \mathcal{V}_2$ or if both $C \in \mathcal{V}_1 \cap \mathcal{V}_2$ and $(M_1, \vec{u}_1) \succeq_C (M_2, \vec{u}_2)$ for all compatible settings $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$, then $\mathcal{F}(C) = \mathcal{F}_1(C)$; if $C \in \mathcal{V}_2 - \mathcal{V}_1$ or if both*

$C \in \mathcal{V}_1 \cap \mathcal{V}_2$ *and* $(M_2, \vec{u}_2) \succeq_C (M_1, \vec{u}_1)$ *for all compatible settings* $(M_1, \vec{u}_1)$ *and* $(M_2, \vec{u}_2)$*, then* $\mathcal{F}(C) = \mathcal{F}_2(C)$.[5] [6]

Note that we assume that when experts use the same variable, they are referring to the same phenomenon. Our approach does not deal with the possibility of two experts using the same variable name to refer to different phenomena.

Returning to Example 3.6, suppose that M_1 and M_2 are compatible. Then $M_1 \oplus M_2$ has the causal graph described in Figure 5; that is, even though **?**] do not have a formal theory for merging models, they actually merge models in just the way that we are suggesting.

The next theorem provides evidence that Definition 3.11 is reasonable and captures our intuitions. To explain the theorem, we introduce a little more notation. We write $M_1 \succeq_C M_2$ if there exist contexts $\vec{u}_1$ and $\vec{u}_2$ such that $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are compatible and $(M_1, \vec{u}_1) \succeq_C (M_2, \vec{u}_2)$. By Lemma 3.8, this is consistent with our definition of $\succeq_C$ in the case of strong domination. We also define $M_1 \sim_C M_2$ if $M_1 \succeq_C M_2$ and $M_2 \succeq_C M_1$, and $M_1 \succ_C M_2$ if $M_1 \succeq_C M_2$ and $M_1 \not\sim_C M_2$. (We use the notation $\succ_C$ in the proof of Theorem 3.13 given in the appendix.) Part (b) says that $\succeq_C$ is well defined, so that in the clauses in the definition where there might be potential conflict, such as in the definition of $\mathcal{F}(C)$ when $C \in \mathcal{V}_1 \cap \mathcal{V}_2$ and $M_1 \sim_C M_2$, there is in fact no conflict; part (a) is a technical result needed to prove part (b). Part (c) says that the merged model is guaranteed to be acyclic. Part (d) says that causal paths in M_1 are preserved in $M_1 \oplus M_2$, while part (e) says that at least as far as formulas involving the variables in M_1 go, $M_1 \oplus M_2$ and M_1 agree. Parts (d) and (e) can be viewed as saying that the essential causal structure of M_1 and M_2 is preserved in $M_1 \oplus M_2$. All conclusions that can be drawn in M_1 and M_2 individually can be drawn in $M_1 \oplus M_2$. (In the language of Halpern [12], part (e) says that $M_1 \oplus M_2$ is essentially a conservative extension of M_1.) But it is important to note that $M_1 \oplus M_2$ lets us go beyond M_1 and M_2, since we can, for example, consider interventions that simultaneously affect variables in M_1 that are not in M_2 and variables in M_2 that are not in M_1. Finally, parts (f) and (g) say that $\oplus$ is commutative and associative over its domain.

Theorem 3.13. *Suppose that* M_1 M_2*, and* M_3 *are pairwise compatible. Then the following conditions hold.*

(a) If $M_1 \sim_C M_2$*, then (i)* $Par_{M_1}(C) = Par_{M_2}(C)$ *and (ii)* $\mathcal{F}_1(C) = \mathcal{F}_2(C)$.

(b) $M_1 \oplus M_2$ *is well defined.*

(c) $M_1 \oplus M_2$ *is acyclic.*

[5] We are abusing notation here and viewing $\mathcal{F}_i(C)$ as a function from the values of the parents of C in M_i to the value of C, as opposed to a function from all the values of all variables other than C to the value of C.

[6] If we allow models where recursivity is context-dependent, then (since Lemma 3.9 no longer holds), we must modify this definition to say that, for all settings $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ and variables C, either $\mathrm{MI1}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ holds or $\mathrm{MI1}_{(M_2,\vec{u}_2),(M_1,\vec{u}_1),C}$ holds (so the direction of domination can depend on the context), and if $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are compatible, then MI4. This results in a somewhat more complicated definition of F_C, where whether M_1 or M_2 is used to define C depends on the context. We omit details here.

(d) *If A and B are variables in M_1, then A is an ancestor of B in M_1 iff A is an ancestor of B in $M_1 \oplus M_2$. If $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are compatible, $\vec{u}$ is a context for $M_1 \oplus M_2$ that agrees with $\vec{u}_1$ on the variables in $\mathcal{U} \cap \mathcal{U}_1$, and φ is a formula that mentions only variables in M_1, then $(M_1, \vec{u}_1) \models \varphi$ iff $(M_1 \oplus M_2, \vec{u}) \models \varphi$.*

(f) *$M_1 \oplus M_2 = M_2 \oplus M_1$.*

(g) *If M_3 is compatible with $M_1 \oplus M_2$ and M_1 is compatible with $M_2 \oplus M_3$, then $M_1 \oplus (M_2 \oplus M_3) = (M_1 \oplus M_2) \oplus M_3$.*

The proof of Theorem 3.13 is rather involved; the details can be found in Appendix A.

We define what it means for a collection $\mathcal{M} = \{M_1, \ldots, M_n\}$ of causal models to be *mutually compatible* by induction on the cardinality of $\mathcal{M}$. If $|\mathcal{M}| = 1$, then mutual compatibility holds by definition. If $|\mathcal{M}| = 2$, then the models in $\mathcal{M}$ are mutually compatible if they are compatible according to Definition 3.11. If $|\mathcal{M}| = n$, then the models in $\mathcal{M}$ are mutually compatible if the models in every subset of $\mathcal{M}$ of cardinality $n-1$ are mutually compatible, and for each model $M \in \mathcal{M}$, M is compatible with $\oplus_{M' \neq M} M'$. By Theorem 3.13, if $M_1, \ldots, M_n$ are mutually compatible, then the causal model $M_1 \oplus \cdots \oplus M_n$ is well defined; we do not have to worry about parenthesization, nor the order in which the settings are combined. Thus, the model $\oplus_{M' \neq M} M'$ considered in the definition is also well defined. Theorem 3.13(e) also tells us that $M_1 \oplus \cdots \oplus M_n$ contains, in a precise sense, at least as much information as each model M_i individually. Thus, by merging mutually compatible models, we are maximizing our use of information.

This approach to merging models is one of the main contributions of this paper. Using it, we show in Section 4 how experts' models can be combined to reason about interventions.

3.2. The complexity of determining compatibility

Checking whether two given causal models M_1 and M_2 are compatible requires checking whether the conditions of Definition 3.11 hold. This amounts to checking the conditions $\mathrm{MI1}_{M_1,M_2,C}$ and $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ for all variables $C \in (\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2)$ and compatible settings $(M, \vec{u}_1)$ and $(M, \vec{u}_2)$.

How hard this is to do depends in part on how the models are presented. If the models are presented *explicitly*, which means that, for each variable C, the equation for C is described as a (huge) table, giving the value of C for each possible setting of all the other variables, the problem is polynomial in the sizes of the input models. However, the size of the model will be exponential in the number of variables.

In this case, checking whether $\mathrm{MI1}_{M_1,M_2,C}$ holds for all C amounts to checking whether the parents of C in M_2 are the immediate M_2-ancestors of C in M_1. To solve this, we need to determine, for each pair of endogenous variables X and Y in M_i for $i = 1, 2$, whether X depends on Y. With this information, we can construct the causal graphs for M_1 and M_2, and then quickly determine whether $\mathrm{MI1}_{M_1,M_2,C}$ holds.

If the model is given explicitly, then determining whether X depends on Y amounts to finding two rows in the table of values of F_X that differ only in the value of Y and in the outcome. As the number of pairs of rows is quadratic in the size of the table, this is polynomial in the size of the input. Thus, we can determine if $\mathrm{MI1}_{M_1,M_2,C}$ holds in polynomial time.

Checking if $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,,\vec{u}_2),C}$ holds amounts to checking whether

$$(M_1, \vec{u}_1) \models [\vec{X} \leftarrow \vec{x}](C = c),$$

iff

$$(M_2, \vec{u}_2) \models [\vec{X} \leftarrow \vec{x}](C = c).$$

For a specific context $\vec{u}$ and choice of $\vec{X}$ and $\vec{x}$, we can easily compute the value of C in a context $\vec{u}$ if $\vec{X}$ is set to $\vec{x}$ (even if the model is not given explicitly). Since the number of possible contexts is smaller than the size of an explicitly presented model, we can also determine whether $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ holds in polynomial time if the model is presented explicitly. Moreover, we can also determine whether $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are compatible in polynomial time.

On the other hand, if the models are presented in a more compact way, using the structural equations, then the (descriptions of the) models are of size polynomial in the number of variables in the model. This makes checking compatibility more difficult, as we now show.

Proposition 3.14. *Given two causal models M_1 and M_2 of size polynomial in the number of variables, determining whether they are compatible is in $\mathrm{P}_{||}^{\mathrm{NP}}$ and is co-NP-hard in the sizes of M_1 and M_2.*

Proof. We prove a slightly stronger claim: that checking $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ is co-NP-complete in the sizes of M_1 and M_2, and that checking $\mathrm{MI1}_{M_1,M_2,C}$ is in $\mathrm{P}_{||}^{\mathrm{NP}}$. The complexity class $\mathrm{P}_{||}^{\mathrm{NP}}$consists of all decision problems that can be solved in polynomial time with parallel (i.e., non-adaptive) queries to an NP oracle (see [27, 19, 20]).

We start by showing that checking that $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ holds is in co-NP by showing that the complementary problem, namely demonstrating that $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ does not hold, is in NP. To do this, we guess a witness: a setting $\vec{x}$ for the common variables $\vec{X}$ of M_1 and M_2 other than C, a value c of C, and contexts $\vec{u}_1$ and $\vec{u}_2$ for M_1 and M_2, respectively, such that $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are compatible,

$$(M_1, \vec{u}_1) \models [\vec{X} \leftarrow \vec{x}](C = c),$$

but

$$(M_2, \vec{u}_2) \not\models [\vec{X} \leftarrow \vec{x}](C = c)$$

(or vice versa). A witness can be verified in polynomial time in the size of the model, as it amounts to assigning values to all variables in the models and checking the value of C.

The proof that the problem is co-NP hard is by reduction from the known co-NP-complete problem *Tautology*: determining whether a Boolean formula φ is a tautology. Let φ be a Boolean formula over the variables $\{Y_1, \ldots, Y_n\}$. We construct a causal model M_1 as follows:

1. $\mathcal{U}_1 = \{U_1, \ldots, U_n\}$;
2. $\mathcal{V}_1 = \{Y_1, \ldots, Y_n, C\}$;
3. $\mathcal{R}_1(X) = \{0, 1\}$ for all $X \in \mathcal{V}_1$;
4. the equations are $Y_i = U_i$ for $i = 1, \ldots, n$ and $C = \varphi$.

In other words, the variables $\{Y_1, \ldots, Y_n\}$ are binary variables in M_1, and the value of C is determined by φ. We note that since the set of exogenous variables is the same in M_1 and in M_2, all causal setting of M_1 and M_2 are compatible.

The second causal model M_2 is constructed as follows:

1. $\mathcal{U}_2 = \{U_1, \ldots, U_n\}$;
2. $\mathcal{V}_2 = \{Y_1, \ldots, Y_n, C\}$;
3. $\mathcal{R}_2(X) = \{0, 1\}$ for all $X \in \mathcal{V}_2$;
4. the equations are $Y_i = U_i$ for $i = 1, \ldots, n$, and $C = 1$.

$\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ holds iff φ is a tautology. Indeed, if φ is a tautology, then $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ holds trivially. On the other hand, if φ is not a tautology, then it is easy to see that $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$ does not hold, since there is some setting of the variables $Y_1, \ldots, Y_n$ that makes $C = 0$.

To prove membership of $\mathrm{MI1}_{M_1,M_2,C}$ in $\mathrm{P}_{||}^{\mathrm{NP}}$, we describe a polynomial-time algorithm for deciding $\mathrm{MI1}_{M_1,M_2,C}$ that makes parallel queries to an NP oracle. We define an oracle $O^{Dep}(M, X, Y)$ as follows: for a causal model M and two variables X and Y of M, it answers "yes" if F_X depends on the variable Y in M and "no" otherwise. It is easy to see that $O^{Dep}(M, X, Y)$ is in NP, since a witness for the positive answer is a pair of assignments to the variables of F_X that differ only in the value of Y and in the result. A witness is clearly verifiable in polynomial time: we simply instantiate F_X on these two assignments and verify that the results are different. (We have implicitly assumed here that the equation F_X can be computed in polynomial time, as it is a part of M.) By querying the oracle $O^{Dep}(M_i, X, Y)$ for all endogenous variables X and Y in M_i, for $i = 1, 2$, we can determine the causal graphs of M_1 and M_2, and thus whether $\mathrm{MI1}_{M_1,M_2,C}$ holds. The number of queries is at most quadratic in the sizes of M_1 and M_2, hence the algorithm terminates in polynomial time. ∎

3.3. BDL's desiderata

We now discuss the extent to which our approach to merging models M_1 and M_2 satisfies BDL's desiderata. Recall that BDL considered only causal networks, not causal models in our sense; they also assume that all models mention the same set of variables. They consider four desiderata. We briefly describe them and their status in our setting.

- *Universal Domain*: The rule for combining models accepts all possible inputs. We weaken this by combining only models that are compatible. We can view compatible models as ones that, in BDL's language, "reflect a certain amount of cohesion across different individuals' causal judgments".

- *Acyclicity*: The result of merging M_1 and M_2 is acyclic. This follows from Theorem 3.13(c), provided that $M_1 \oplus M_2$ is defined.

- *Unbiasedness*: if $M_1 \oplus M_2$ is defined, and M_1 and M_2 mention the same variables, then whether B is a parent of C in $M_1 \oplus M_2$ depends only on whether B is a parent of C in M_1 and in M_2. This property holds trivially for us, since if B and C are in both M_1 and M_2 and $M_1 \oplus M_2$ is defined, then B is a parent of C in $M_1 \oplus M_2$ iff B is a parent of C in both M_1 and M_2. (The version of this requirement given

by BDL does not say "if $M_1 \oplus M_2$ is defined", since they assume that arbitrary models can be merged.)

BDL also have a *neutrality* requirement as part of unbiasedness. Unfortunately, an aggregation rule that says that B is a parent of C in $M_1 \oplus M_2$ iff B is a parent of C in both M_1 and M_2 (which seems quite reasonable to us) is not neutral in their sense. That is because it follows from the BDL formal definition of neutrality that if M_1 says that B is a parent of C and M_2 says that B is not a parent of C, then B is a parent of C in $M_1 \oplus M_2$ iff B is not a parent of C in $M_2 \oplus M_1$. So, a consequence of their definition is that $\oplus$ cannot be commutative (since we cannot have $M_1 \oplus M_2 = M_2 \oplus M_1$ if B is a parent of C in M_1 but not in M_2). By way of contrast, in our definition of $\oplus$, if B is a parent of C in M_1 but not in M_2, and M_1 and M_2 are compatible, then B is a parent of C in neither $M_1 \oplus M_2$ nor $M_2 \oplus M_1$ (and $M_1 \oplus M_2 = M_2 \oplus M_1$). In light of this observation, we do not consider neutrality a reasonable requirement to satisfy.

- *Non-dictatorship:* no single expert determines the aggregation. This clearly holds for us.

3.4. Decomposition of causal models

While the notion of dominance used in Definition 3.11 is useful, it still does not cover many cases of interest. The following example considers causal models for the emergence of radicalization in US prisons. The material is taken from **?**]. Although Useem and Clayton do not provide causal models, we construct these based on the description provided. Below we provide a detailed explanation of all the variables and their dependencies.

Example 3.15. Consider the two causal models in Figure 6. M_1 represents Expert 1's opinion about the causes of emergence of a radicalizing setting (R) in the State Correctional Institution Camp Hill in Pennsylvania. M_2 represents Expert 2's opinion about the causes of emergence of a radicalizing setting in the Texas Department of Corrections and Rehabilitation. Both experts agree on the structural equations for R, that is, the emergence in both prison settings is attributed to the same three factors: "order in prisons" (PD), "a boundary between the prison and potentially radicalizing communities" (CB), and "having missionary leadership within the prison organizations" (AM). They also both share the same outcome: the emergence of a radicalizing setting (R). However, the experts differ on the structural equations for PD, CB and AM. As can be observed from the descriptions provided, some variables and their dependency relations are specific to a prison. In M_1, PD is attributed to corruption (CG) and lax management (LM) in the prison's staff together with prisoners being allowed to roam freely (FM). CB is viewed as a result of religious leaders within the facilities being permitted to provide religious services freely (IL) and by prisoners showing a form of membership within a prison community (CM); the latter is signalled by prisoners being allowed to wear distinguished street clothing (SC). Prison authorities' exercizing of internal punishments, such as administrative segregation (AS), away from external oversight, and IL are considered to directly contribute to AM. M_2 instead considers PD to be linked to the rapid growth in inmate numbers (RG), inmates being allowed to assist authorities in maintaining order (AA), and inmates feeling significantly deprived (D) within the prisons—the latter

as a result of being forced to engage in unpaid work (W) and having limited contact with visitors (C). We can show that $\text{MI1}_{M_1,M_2,CB}$ and $\text{MI1}_{M_1,M_2,AM}$ hold. However, neither $\text{MI1}_{M_1,M_2,PD}$ nor $\text{MI1}_{M_2,M_1,PD}$ holds. Therefore the models are not compatible according to Definition 3.11. ∎

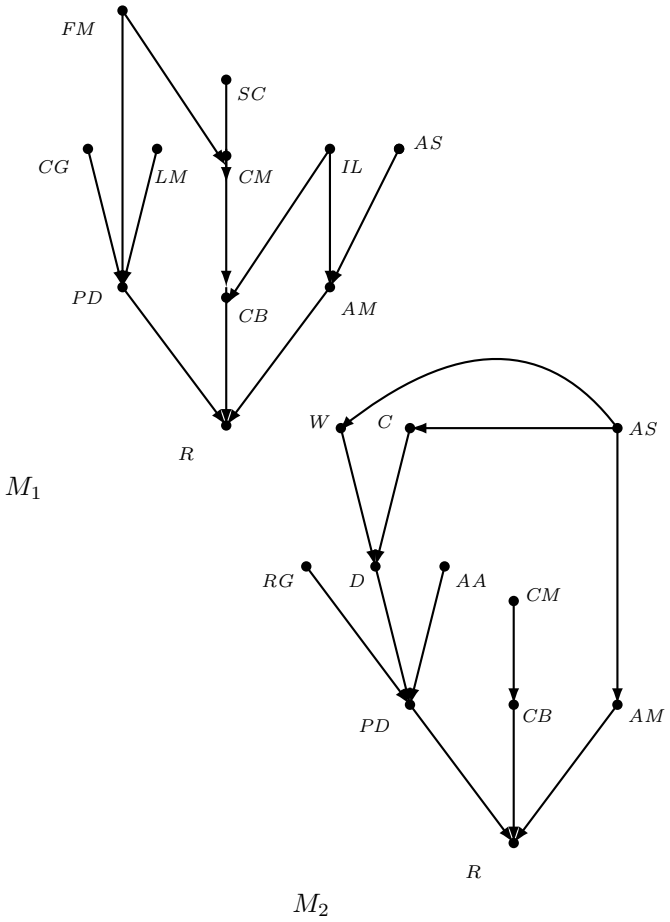


Figure 6: Schematic representation of the two prison models.

The example above illustrates that two experts' models might not be compatible. But we would expect that these models have submodels that are compatible. Finding such submodels has several advantages. First, consider the situation where the policymaker is given several different causal models that are not compatible according to Definition 3.11. If we could decompose the models, we might be able to "localize" the incompatibility, and merge the parts of the models that are compatible. Doing so may suggest effective interventions. Another advantage of decomposing a model is that it allows the policymaker

to reason about each submodel in isolation. Since the problem of computing causes is DP-complete and the problem of computing interventions is co-NP-complete, having a smaller model to reason about could have a significant impact on the complexity of the problem.

In order to define the notion of decomposition, we need some preliminary definitions.

Definition 3.16. *[Order-preserving partition] A sequence $\langle \mathcal{V}_1, \ldots, \mathcal{V}_k \rangle$ of subsets of variables in $\mathcal{V}$ variables in a causal model M is an* order-preserving partition *if $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$ for $i \neq j$, $\cup_{i=1}^{k} \mathcal{V}_i = \mathcal{V}$ (so $\{\mathcal{V}_1, \ldots, \mathcal{V}_k\}$ is a partition of $\mathcal{V}$), and for all i, j with $i < j$, no variable in $\mathcal{V}_j$ is an ancestor of a variable in $\mathcal{V}_i$.*

Definition 3.17. *[Decomposable causal models] $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$ is* decomposable *if there exist $k \geq 1$ mutually compatible causal models $\{M_i = ((\mathcal{U}_i, \mathcal{V}_i, \mathcal{R}_i), \mathcal{F}_i) : 1 \leq i \leq k\}$ such that $\langle \mathcal{V}_1, \ldots, \mathcal{V}_k \rangle$ is an order-preserving partition of $\mathcal{V}$, $\mathcal{U}_i$ consists of all the endogenous and exogenous variables in M not in $\mathcal{V}_i$ that are parents of some variable in $\mathcal{V}_i$ in M, and $\mathcal{F}_i$ is just the restriction of $\mathcal{F}$ to the variables in $\mathcal{V}_i$. $M_1, \ldots, M_k$ is called a* decomposition *of M.*

Lemma 3.18. *If $M_1, \ldots, M_k$ is a decomposition of M, then $M_1 \oplus \cdots \oplus M_k = M$.*

Proof. The proof is immediate given the observation that we do not change any of the structural equations of M when decomposing it into submodels. ∎

It is easy to see that, for a given model, there can be many ways to decompose it into a set of submodels according to Definition 3.17. Moreover, all models are decomposable by Definition 3.17. Indeed, any model M can be trivially decomposed to $|\mathcal{V}|$ submodels, each of which consists of exactly one endogenous variable. Of course, such a decomposition is useless for practical purposes; the decompositions we consider are those that help in either analyzing the model or reducing the complexity of computing causes. Note that, while the set $\mathcal{U}_i$ of exogenous variables in a component M_i of a decomposition is a superset of $\mathcal{U}$, the equations are identical to those of the original model, so M_i is in fact simpler than the original model (and possibly much simpler). We expect that, in practice, decomposing a model will make computations far simpler. In Example 3.19 below, we illustrate a nontrivial decomposition.

Example 3.19. Consider the causal models in Figure 6 from the prison example (Example 3.15). Let $\mathcal{U}_1$ and $\mathcal{U}_2$ be the set of exogenous variables for M_1 and M_2, respectively (which are not explicitly given in Figure 6). By Definition 3.17, we can decompose M_1, into M_{11}, M_{12}, M_{13}, and M_{14}, where $M_{1j} = ((\mathcal{U}_{1j}, \mathcal{V}_{1j}, \mathcal{R}_{1j}), \mathcal{F}_{1j})$, $\mathcal{V}_{11} = \{FM, CG, LM, PD\}$, $\mathcal{U}_{11}$ consists of all the exogenous variables in $\mathcal{U}_1$ that are ancestors of the variables in $\mathcal{V}_{11}$, $\mathcal{V}_{12} = \{SC, CM, CB, IL\}$, $\mathcal{U}_{12}$ consists of all the exogenous variables in $\mathcal{U}_1$ that are ancestors of the variables in $\mathcal{V}_{12}$ together with FM, $\mathcal{V}_{13} = \{AS, AM\}$, $\mathcal{U}_{13}$ consists of all the exogenous variables in $\mathcal{U}_1$ that are ancestors of $\mathcal{V}_{13}$ together with IL, $\mathcal{V}_{14} = \{R\}$, and $\mathcal{U}_{14} = \{PD, CB, AM\}$. Similarly we can decompose M_2 into four submodels M_{21}, M_{22}, M_{23}, and M_{24}, where $M_{2j} = ((\mathcal{U}_{2j}, \mathcal{V}_{2j}, \mathcal{R}_{2j}), \mathcal{F}_{2j})$, $\mathcal{V}_{21} = \{C, W, AA, D, RG, PD\}$, $\mathcal{U}_{21}$ consists of all variables of $\mathcal{U}_2$ that are ancestors of the variables in $\mathcal{V}_{21}$ together with AS, $\mathcal{V}_{22} = \{CM, CB\}$, $\mathcal{U}_{22}$ consists of all the variables in $\mathcal{U}_2$ that are ancestors of the variables in $\mathcal{V}_{22}$, $\mathcal{V}_{23} = \{AS, AM\}$, $\mathcal{U}_{23}$ consists of all the variables in $\mathcal{U}_2$ that are ancestors of the variables in $\mathcal{V}_{23}$, $\mathcal{V}_{24} = \{R\}$, and $\mathcal{U}_{24} = \{PD, CB, AM\}$. Figures 7 and 8 show the

four submodels resulting from these decompositions (with the exogenous variables that are in $\mathcal{U}_1$ and $\mathcal{U}_2$ omitted). There is some flexibility in how we do the decomposition. For example, we could move AS from $\mathcal{V}_{23}$ to $\mathcal{V}_{21}$. We would then need to remove AS from $\mathcal{U}_{21}$ and add the parents of AS to $\mathcal{U}_{21}$. Then in M_{23} we would remove AS from $\mathcal{V}_{23}$; AS would be an exogenous parent of AM. In addition, we would remove the parents of AS from $\mathcal{U}_{23}$ (unless they were also parents of AM). However, we cannot, for example, move CB from V_{i2} to $\mathcal{V}_{i1}$, as then $\langle \mathcal{V}_{i1}, \mathcal{V}_{i2}, \mathcal{V}_{i3}, \mathcal{V}_{i4} \rangle$ would not be an order-preserving partition (since FM is an ancestor of CM, which is an ancestor of CB). ∎

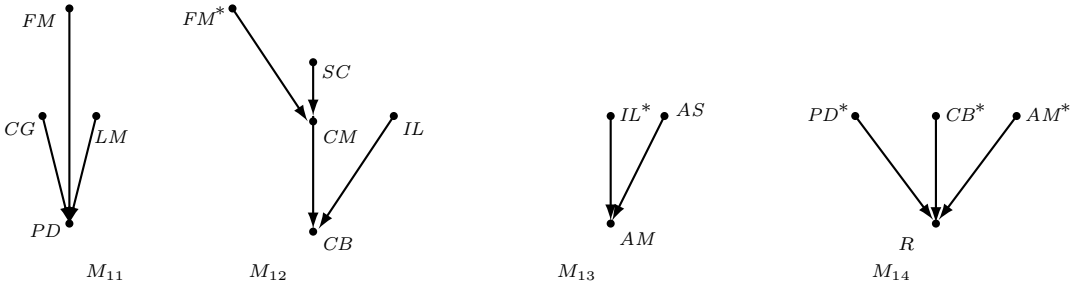


Figure 7: Decomposition of the model M_1 from Example 3.15. We label variables in $\mathcal{V}_1$ that are exogenous in submodel M_{1j} with an asterisk.

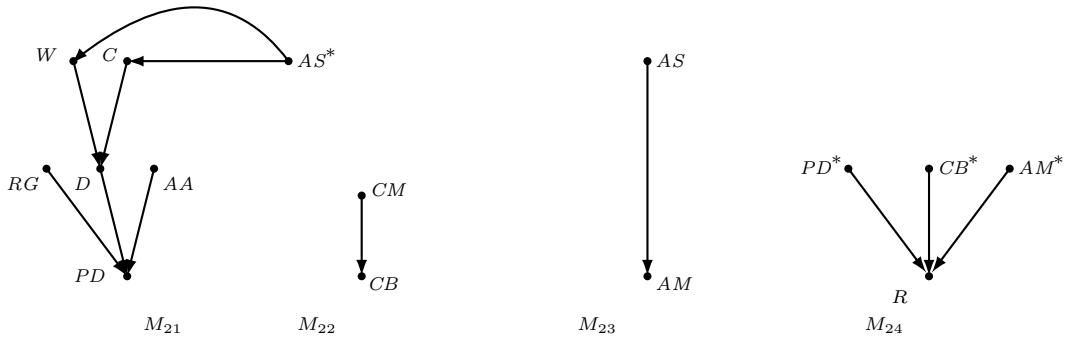


Figure 8: Decomposition of the model M_2 from Example 3.15. We label variables in $\mathcal{V}_2$ that are exogenous in submodel M_{2j} with an asterisk.

Decomposing incompatible models into smaller submodels can in some cases help determine common interventions over shared outcomes in the original models despite their incompatibility. Consider, for example, the two models M_1 and M_2 in the prison

example. Although they are incompatible (as observed in Example 3.15), the submodels M_{12} and M_{22} in Figures 7 and 8, respectively, obtained from their decomposition, are compatible. We have $M_{12} \succeq_{CB} M_{22}$ and $M_{12} \succeq_{CM} M_{22}$. The composition of the two submodels yields a merged model equivalent to M_{12}. Given this, it may be concluded that interventions over *SC* or *IL* make it possible to change the value of *CB* in the two models M_1 and M_2 and ultimately *R* (assuming that both models M_1 and M_2 share the structural equation $R = PD \wedge CB \wedge AM$). Note, however, that in the decomposition illustrated above, M_{11} and M_{21} are incompatible, since we have neither $M_{11} \not\succeq_{PD} M_{12}$ nor $M_{12} \not\succeq_{PD} M_{11}$. Therefore, we cannot determine the effect of interventions on *PD*.

Another advantage of decomposing a causal model M into a set of smaller submodels is that we can reason about each submodel separately. In particular, we can compute the set of causes and possible interventions for a given outcome. However, in order to use these results to reason about the whole model, we need to perform additional calculations. Informally, when decomposing M into a set of smaller submodels, we can view each submodel as a black box, with inputs and outputs being the exogenous variables of the submodel and the leaves in the causal graph of the submodel, respectively. We can then connect these variables into an abstract causal graph for the original model, essentially ignoring the internal variables. If the submodels are fairly large, the graph of submodels will be significantly smaller than the causal network of M. We can then apply causal reasoning to the abstract graph, which will result in a set of submodels being causes for the outcome. For these submodels, we can calculate the causes of their outcomes for each submodel separately. As causality is *DP*-complete, and computing interventions is co-*NP*-complete, solving a set of smaller problems instead of a large problem is cheaper.

We note that, in fact, interesting decompositions (that is, decompositions of a large model into a set of submodels of reasonable sizes with relatively few interconnections between them, which means that we can analyze causality both within a submodel and between submodels relatively easily) are possible only in models that are somewhat loosely connected. Such a decomposition can often be done for real-life cases; see Example 3.20. We believe that, in practice, analyzing the effect of interventions in a model will be difficult precisely when a model is highly connected, so that there are many causal paths. We expect the causal models that arise in practice to be much more loosely connected, and thus amenable to useful decompositions. Hence, the computation of causes and interventions in practice should not be as bad as what is suggested by our worst-case analysis.

Below, we briefly discuss the relevant aspects of two cases of child abuse that resulted in the death of a child: the "Baby P" case and the Victoria Climbiè case. In these cases, experts' opinions were in fact not compatible, and there were natural ways to decompose the causal models.

Example 3.20 (The cases of Baby P and Victoria Climbiè)**.** Baby P (Peter Connelly) died in 2007 after suffering physical abuse over an extended period of time [**?**]. The court ultimately found the three adults living in a home with baby Peter guilty of "causing or allowing [Peter's] death" [**?**]. After baby Peter's death, there was an extensive inquiry into practices, training, and governance in each of the involved professionals and organizations separately.[7]

[7] Chockler et al. [2] provide a more detailed discussion of the case of "Baby P", including a construction

As shown by Chockler et al. [2], the complete causal model for the Baby P case is complex, involving many variables and interactions between them. There were several authorities involved in the legal proceedings, specifically social services, the police, the medical system, and the court. In addition, the legal proceedings considered the family situation of Baby P. Roughly speaking, the causal model can be viewed as having the schematic breakdown presented in Figure 9.

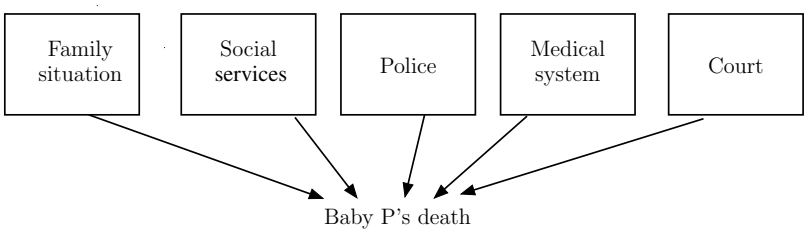


Figure 9: Schematic representation of causal submodels in the Baby P case.

Each of the experts involved in the legal inquest and enquiry had expertise that corresponded to one of the boxes in Figure 9 (i.e., there were no experts with expertise that covered more than one box). The figure suggests that we might divide the causal model into submodels corresponding to each box. The schematic representation in Figure 9 does not take into account the interaction between submodels. In reality, there were numerous interactions between, for example, the social services and the court submodels, leading to court hearings, which in turn determined the course of action taken by the social services and the police after the court decision. Once we model these interactions more carefully, we need a somewhat more refined decomposition.

We give a decomposition in Figure 10 that takes into account the interactions for part of the case, namely, the part that concerns the social services, the court, the police, and family life. To make the decomposition consistent with Definition 3.17, we break up social services into two submodels, for reasons explained below.

The variables in the figure are: FV for whether there was a family visit from the social services; PR for whether there was a police report; CH for whether there was a court hearing; RFH for whether the child was removed from home; CP for whether the child was put on the Child Protection Register; SR for whether there was a social services report; CS for whether the child was declared safe in the family home; MA, PA, and OA for whether the child was abused by his mother, the mother's partner, or another adult in the house, respectively; CA for whether the child was abused; and, finally, D for whether the final outcome was death (of Baby P) due to abuse. Note that, as usual, we have omitted exogenous variables of the full model in the figure; it shows only the endogenous variables. Thus, we do not have the exogenous variables that determine SR, FV, PR, MA, PA, or OA. The dotted rectangles in Figure 10 determine a decomposition. Each rectangle consists of the endogenous variables of one submodel. The exogenous variables of the SocialServices#2 and Outcome submodels are those parent variables appearing in the other submodels. Thus, for example, in the Outcome submodel, the exogenous

of the causal model.

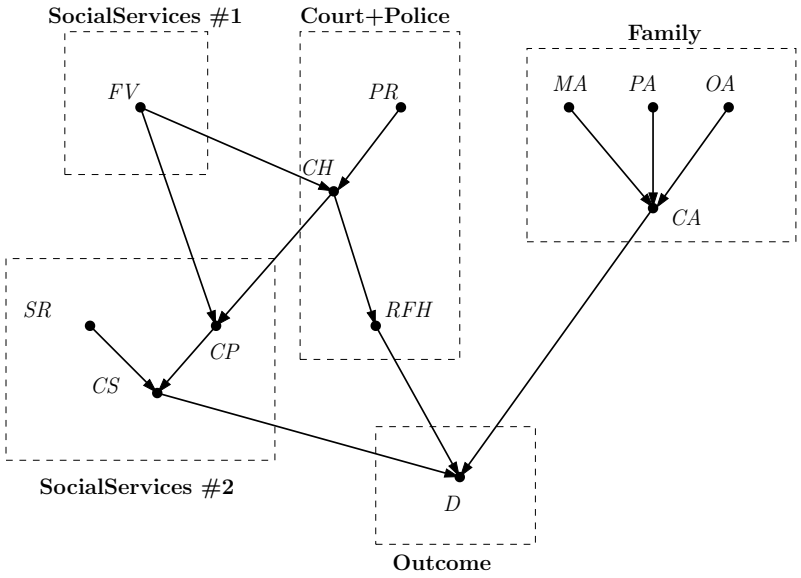


Figure 10: Simplified causal model M of a part of the Baby P case.

variables are CS, RFH, and CA. The submodels are described in Figure 11. The dotted rectangles in Figure 10 can be viewed as compact representations of the submodels in Figure 11.

Of course, there is more than one way to decompose the model of Figure 10. For example, the submodel currently standing for the court and the police can be decomposed into two smaller submodels, one for the court and one for the police. However, it is critical that social services is decomposed into two submodels. The variable CH depends on FV, and the variable CP in turn depends on CH, hence FV and CP cannot be in the same submodel (or else we would violate the requirement of Definition 3.17 that the sets of endogenous variables of each submodel form an order-preserving partition of the endogenous variables of the original model).

We consider another case of child abuse that resulted in child's death: Victoria Climbiè [**?**]. Victoria died in her house from hypothermia in February 2000, 18 months after arriving in the UK from the Ivory Coast to live with her great-aunt. Her great-aunt and the great-aunt's boyfriend were found guilty of Victoria's murder (in contrast with the Baby P case, where the adults in the house were found guilty of causing or allowing his death).

The inquiry into the circumstances of Victoria's death placed the blame on social workers, who failed to notice Victoria's injuries, paediatricians, who accepted the explanation of Victoria's great-aunt that Victoria's injuries were self-inflicted, and the metropolitan police. In addition, the inquiry noted that the pastors in the church to which Victoria's great-aunt belonged, had concerns about the child's well-being but failed to contact any child protection services. The inquiry suggested several interventions on the procedures of social workers and paediatricians. These interventions turned out to be inadequate, as the death of Baby P occurred seven years later under somewhat similar

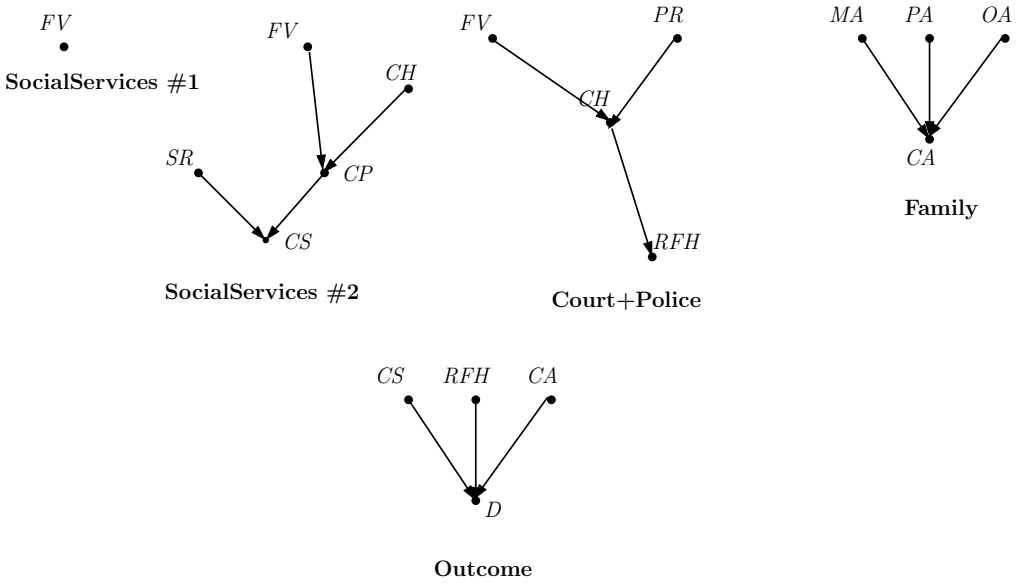


Figure 11: One possible decomposition of the Baby P model M.

circumstances, and the abuse also went unnoticed until his death.

Although there were some similarities between the Baby P case and the Victoria Climbiè cases, there were also some differences. For example, while Vicitoria Climbiè died at home, Baby P died in the hospital. Thus, the causal models for these two cases differ somewhat. However, the causal model for the Victoria Climbié case is also decomposable into compatible submodels in the sense of Section 3.4. Moreover, some of the submodels in the decomposition are identical to those in the causal model for Baby P. Specifically, there are submodels for the police, the medical system, the family system, and the courts, just as in the case of Baby P, as well as a submodel for the church. The schematic breakdown is presented in Figure 12. Although we do not provide the causal model in

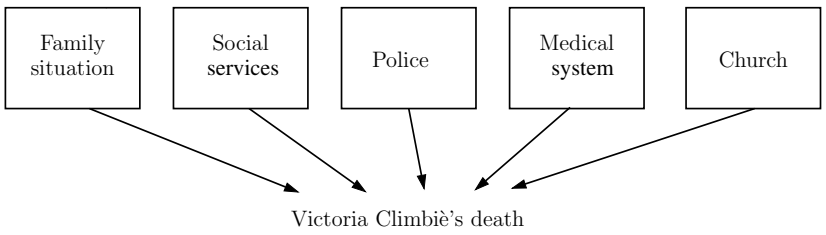


Figure 12: Schematic representation of causal submodels in the Victoria Climbiè's case.

detail here, this discussion already illustrates a major advantage of decomposition: it allows us to reuse causal models that were developed in one case and apply them to another, thus saving a lot of effort. Moreover, if the same submodel appears in several different cases, such as the social services submodel in these examples, this suggests that

the policymaker should prefer interventions that address the problems demonstrated by this submodel, as they are likely to affect several cases. In fact, the cases of child abuse that remains undetected due to problems in the social services sadly continue to occur (see, for example, the recently published cases discussed in [**?**]). Even though the causal models for different cases will undoubtedly be different, we can still take advantage of the common submodels. We expect that this will be the case in many other situations as well. ▮

From a practical perspective, Example 3.20 demonstrates one benefit of decomposition: the decomposition allows us to capture different aspects of the case, each requiring different expertise. This facilitates different experts working on each of the submodels independently. The process also works in the other direction: a policymaker often has a crude idea of the general structure of the causal model, and what components are involved in the decision-making process. She can then decompose her initial causal model into submodels and, guided by these submodels, decide which areas of expertise are critical.

A further benefit of decomposition illustrated by these examples is that, although different, the causal models had some common submodels. Thus, decomposition supports a form of modularity in the analysis, and enables results of earlier analyses to be reused.

4. Combining Experts' Opinions

In this section, we show how we can combine experts' causal opinions. Suppose that we have a collection of pairs (M_1, p_1), ..., (M_n, p_n), with $p_i \in (0, 1]$; we can think of M_i as the model that expert i thinks is the right one and p_i as the policymaker's prior degree of confidence that expert i is correct. (The reason we say "prior" here will be clear shortly.) Our goal is to combine the expert' models. We present one way of doing so, that uses relatively standard techniques. The idea is to treat the probabilities $p_1, \ldots, p_n$ as mutually independent. In other words, the policymaker's confidence in the correctness of expert i is independent of her confidence in the correctness of expert j, for $1 \leq i \neq j \leq n$. Thus, if I is a subset of $\{1, \ldots, n\}$, the prior probability that exactly the experts in I are right, which we denote p_I, is $p_I = \prod_{i \in I}(p_i) \cdot \prod_{j \notin I}(1 - p_j)$. Now we have some information regarding whether all the experts in I are right. Specifically, if the models in $\{M_i : i \in I\}$ are not mutually compatible, then it is impossible that all the experts in I are right. Intuitively, we want to condition on this information. We proceed as follows.

Let $\mathit{Compat} = \{I \subseteq \{1, \ldots, n\} : \text{the models in } \{M_i : i \in I\} \text{ are mutually compatible}\}$. For $I \in \mathit{Compat}$, define $M_I = \oplus_{i \in I} M_i$. By Proposition 3.13, M_I is well defined. The models in $\mathcal{M}_{\mathit{Compat}} = \{M_I : I \in \mathit{Compat}\}$ are the candidate merged models that the policymaker should consider. M_I is the "right" model provided that exactly the experts in I are right. But even if $M_I \in \mathcal{M}_{\mathit{Compat}}$, it may not be the "right" model, since it may be the case that not all the expert in I are right. The probability that the policymaker should give M_I is p_I/N, where $N = \sum_{I \in \mathit{Compat}} p_I$ is a normalization factor.

This approach gives the policymaker a distribution over causal models. This can be used to compute, for each context, which interventions affect the outcome φ of interest, and then compute the probability that a particular intervention is effective (which can be done summing the probability of the models M_I in $\mathcal{M}_{\mathit{Compat}}$ where it is effective, which

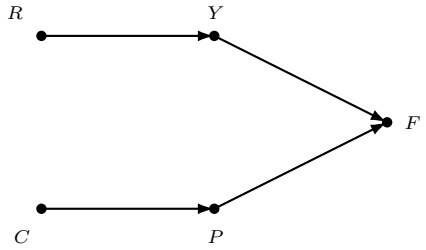


Figure 13: Third expert's (and merged) model of famine.

in turn can be computed as described in Section 3). Note that our calculation implicitly conditions on the fact that at least one expert is right, but allows for the possibility that only some subset of the experts in I is right even if $I \in \mathit{Compat}$; we place positive probability on $M_{I'}$ even if I' is a strict subset of some $I \in \mathit{Compat}$. This method of combining experts' judgments is similar in spirit to the method proposed by Dawid [5] and Fenton et al. [7].

To get a sense of how this works, consider a variant of Example 3.5, in which a third expert provides her view on causes on famine and thinks that government corruption is an indirect cause via its effect on political conflict (see Figure 13); call this model M_3. For simplicity, we assume that all models have the same set of exogenous variables. According to the compatibility definition in Section **??**, the models M_2 and M_3 are compatible (assuming that MI3 holds), but M_1 and M_3 are not. We have $\mathcal{M}_{\mathit{Compat}} = \{\{M_1\}, \{M_2\}, \{M_3\}, \{M_{2,3}\}\}$ with $M_{2,3} = M_2 \oplus M_3 = M_3$. Suppose that experts are assigned the confidence values as follows: $(M_1, 0.4)$, $(M_2, 0.6)$ and $(M_3, 0.5)$ respectively. Then the probability on $M_{2,3}$ is the probability of M_2 and M_3 being right (i.e., $0.6 * 0.5$) and M_1 being wrong (i.e., $1 - 0.4 = 0.6$). So we have

$$\begin{array}{lcl} p_1 & = & 0.4 * 0.4 * 0.5/0.56 = 0.14 \\ p_2 & = & 0.6 * 0.6 * 0.5/0.56 = 0.32 \\ p_3 & = & 0.6 * 0.4 * 0.5/0.56 = 0.21 \\ p_{2,3} & = & 0.6 * 0.5 * 0.6/0.56 = 0.32 \end{array}$$

where $0.08 + 0.18 + 0.12 + 0.18 = 0.56$ is the normalization factor N.

Let us consider the Sampson's domestic violence models as another point of illustration. The model shown in Figure 5 is the result of merging the two compatible models given in Figure 4. We thus have $\mathcal{M}_{\mathit{Compat}} = \{\{M_1\}, \{M_2\}, \{M_{1,2}\}\}$ with $M_{1,2} = M_1 \oplus M_2$ as given in Figure 5. Assuming that expert 1 is assigned a confidence value 0.6 and expert 2 is assigned 0.7, then we have

$$\begin{array}{lcl} p_1 & = & 0.6 * 0.3 * 0.58/0.44 = 0.23 \\ p_2 & = & 0.4 * 0.7 * 0.58/0.44 = 0.36 \\ p_{1,2} & = & 0.6 * 0.7 * 0.42/0.44 = 0.41 \end{array}$$

Note that the number of models in $\mathcal{M}_{\mathit{Compat}}$ may be exponential in the number of experts. For example, if the experts' models are mutually compatible, then *Compat* consists of all subsets of $\{1, \ldots, n\}$. The straightforward computation of interventions per model is exponential in the number of variables in the model. Since the number of variables in a merged model is at most the sum of the variables in each one, the problem

is exponential in the number of experts and the total number of variables in the experts' models. In practice, however, we do not expect this to pose a problem. For the problems we are interested in, there are typically few experts involved; moreover, as we argued in Section 3, policymakers, in practice, restrict their attention to interventions on a small set of variables. Thus, we expect that the computation involved to be manageable.

Up to now, we have assumed that each expert proposes only one deterministic causal model. An expert uncertain about the model can propose several (typically incompatible) models, with a probability distribution on them. We can easily extend our framework to handle this.

Suppose that expert i, with probability p_i of being correct, proposes m models $M_{i1}, \ldots, M_{im}$, where model M_{ij} has probability q_j of being the right one, according to i. To handle this, we simply replace expert i by m experts, $i_1, \ldots, i_m$, where expert i_j proposes model M_{ij} with probability $p_i q_j$ of being correct. As long as each of a few experts has a probability on only a few models, this will continue to be tractable.

5. Conclusions

We have provided a method for merging causal models whenever possible, and used that as a basis for combining experts' causal judgments in a way that gets around the impossibility result of Bradley et al. [1]. We provided a formal definition of compatibility and determined the complexity of checking the compatibility of models. We also presented a notion of model decomposition that allows us to merge submodels from incomptabile models. Our approach can be viewed as a formalization of what was done informally in earlier work [2**?**]. While our requirements for compatibility are certainly nontrivial, the examples that we have considered do suggest that our approach is quite applicable. That said, it would be interesting to consider alternative approaches to combining experts' opinions. The approach considered by Friedenberg and Halpern [8] is one such approach; there may well be others.

In any case, we believe that using causal models as a way of formalizing experts' judgments, and then providing a technique for combining these judgments, will prove to be a powerful tool with which to approach the problem of finding the best intervention(s) that can be performed to ameliorate a situation.

Appendix A. Proof of Theorem 3.13

Proof.

For part (a), suppose that $M_1 \sim_C M_2$, but $Par_{M_1}(C) \neq Par_{M_2}(C)$. We can assume without loss of generality that there is some variable $Y \in Par_{M_1}(C) - Par_{M_2}(C)$. Let $\vec{Z} = Par_{M_1}(C) - \{Y\}$. Since Y is a parent of C in M_1, there must be some setting $\vec{z}$ of the variables in $\vec{Z}$ and values y and y' for Y such that $F^1_C(y, \vec{z}) \neq F^1_C(y', \vec{z})$ in M_1, where $F^1_C = \mathcal{F}_1(C)$. Suppose that $F^1_C(y, \vec{z}) = c$ and $F^1_C(y', \vec{z}) = c'$. Let $\vec{X} = (\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2)$. Since $\mathrm{MI1}_{M_1, M_2, C}$ must hold, it follows that $(Par_{M_1}(C) \cup Par_{M_2}(C)) \subseteq \vec{X}$. From the definition of $\succeq_C$, since $M_1 \sim_C M_2$, there must exist contexts $\vec{u}_1$ and $\vec{u}_2$ such that $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are compatible. Let $\vec{x}$ be a setting of the variables in $\vec{X} - \{C\}$ such that $\vec{x}$ agrees with $\vec{z}$ for the variables in $\vec{Z}$ and $\vec{x}$ assigns y to Y. Let $\vec{x}'$ be identical to $\vec{x}$ except that it assigns y' to Y. Since the values of the variables in $Par_{M_1}(C)$ determine

the value of C in M_1, we have $(M_1, \vec{u}_1) \models [\vec{X} \leftarrow \vec{x}](C = c)$ and $(M_1, \vec{u}_1) \models [\vec{X} \leftarrow \vec{x}'](C = c')$. Since $\vec{x}$ and $\vec{x}'$ assign the same values to all the variables in $Par_2(C)$, we must have $(M_2, \vec{u}_2) \models [\vec{X} \leftarrow \vec{x}](C = c)$ iff $(M_2, \vec{u}_2) \models [\vec{X} \leftarrow \vec{x}'](C = c)$. Thus, we get a contradiction to $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$. It follows that $Par_{M_1}(C) = Par_{M_2}(C)$. The fact that $\mathcal{F}_1(C) = \mathcal{F}_2(C)$ also follows from $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$. For suppose that $\vec{z}$ is a setting of the variables in $Par_1(C) = Par_2(C)$ and $\vec{x}$ is a setting of the variables in $\vec{X}' = \vec{X} - \{C\}$ that agrees with $\vec{z}$ on the variables in $Par_1(C)$. Then we have that $F^1_C(\vec{z}) = c$ iff $(M_1, \vec{u}_1) \models [\vec{X}' \leftarrow \vec{x}](C = c)$ iff $(M_2, \vec{u}_2) \models [\vec{X}' \leftarrow \vec{x}](C = c)$ (by $\mathrm{MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),C}$) iff $F^2_C(\vec{z}) = c$. Thus, $\mathcal{F}_1(C) = \mathcal{F}_2(C)$.

For part (b), note that $M_1 \oplus M_2$ is well defined unless (i) $\mathcal{R}_1(C) \neq \mathcal{R}_2(C)$ for some $C \in ((\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2))$ or (ii) for some $C \in \mathcal{V}_1 \cap \mathcal{V}_2$, we have that either $\mathcal{R}_1(C) \neq \mathcal{R}_2(C)$ or $M_1 \sim_C M_2$ but $\mathcal{F}_1(C) \neq \mathcal{F}_2(C)$. Since M_1 and M_2 are compatible, (i) cannot happen; by part (a), (ii) cannot happen.

For part (c), we first show part (d): if A and B are both nodes in M_1 (i.e., A and B are in $\mathcal{U}_1 \cup \mathcal{V}_1$), then (the node labeled) A is an ancestor of (the node labeled) B in (the causal graph corresponding to) M_1 iff A is an ancestor of B in $M_1 \oplus M_2$, and similarly for M_2.

Suppose that A is an ancestor of B in M_1. Then there is a finite path $A_0, \ldots, A_n$ in the causal graph for M_1, where $A_0 = A$ and $A_n = B$. We first show that if $A_0, \ldots, A_n$ is an arbitrary sequence of nodes in M_1 such that none of the intermediate nodes (i.e., $A_1, \ldots, A_{n-1}$) is in M_2, and either $A_0 = A_n$ or at most one of A_0 and A_n is in M_2, then $A_0, \ldots, A_n$ is a path in M_1 iff $A_0, \ldots, A_n$ is a path in $M_1 \oplus M_2$. We proceed by induction on n, the length of the path. Since all the nodes in M_1 are nodes in $M_1 \oplus M_2$, the result clearly holds if $n = 0$. Suppose that $n > 0$ and the result holds for $n-1$; we prove it for n. If $A_n \in \mathcal{U}_1 - \mathcal{V}_2$, then A_n has no parents in M_1 or $M_1 \oplus M_2$, so the result holds trivially: there is no path $A_0, \ldots, A_n$ in either M_1 or $M_1 \oplus M_2$. If $A_n \in \mathcal{U}_1 \cap \mathcal{V}_2$, then A_n is in M_2 and its parents in $M_1 \oplus M_2$, if it has any, must also be in M_2. Hence, by assumption, none of its parents in $M_1 \oplus M_2$ can be among $A_0, \ldots, A_{n-1}$, and again, the result holds trivially. If $A_n \in \mathcal{V}_1$ and $M_1 \succeq_{A_n} M_2$ (recall that this means that $(M_1, \vec{u}_1) \succeq_{A_n} (M_2, \vec{u}_2)$ for some compatible settings $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$), then $\mathcal{F}_{1,2}(A_n) = \mathcal{F}_1(A_n)$, so the parents of A_n in M_1 are also the parents of A_n in $M_1 \oplus M_2$. In particular, A_{n-1} is a parent of A_n in $M_1 \oplus M_2$ iff A_{n-1} is a parent of A_n in $M_1 \oplus M_2$, and the result follows from the induction hypothesis. Finally, if $A_n \in \mathcal{V}_1 \cap \mathcal{V}_2$ and $M_2 \succeq_{A_n} M_1$, then $\mathcal{F}_{1,2}(A_n) = \mathcal{F}_2(A_n)$, so again, all of the parents of A_2 in $M_1 \oplus M_2$ must be in M_2, and the result holds trivially.

Now suppose that there are $m > 0$ nodes in M_2 on the path from A to B in M_1, say $C_1, \ldots, C_m$, in that order. We show that (i) C_m is an ancestor of B in $M_1 \oplus M_2$, (ii) A is an ancestor of C_1 in $M_1 \oplus M_2$, and (iii) C_1 is an ancestor of C_m in $M_1 \oplus M_2$. Parts (i) and (ii) follow from the earlier argument, since there are no intermediate nodes in M_2 on the path from C_m to B or on the path from A to C_1. So it remains to prove part (iii). We proceed by induction on m. If $m = 1$, the result is trivially true, since C_1 is a node in $M_1 \oplus M_2$. So suppose that $m > 1$. Since M_1 and M_2 are compatible and C_2 is a node in both M_1 and M_2 for $j > 1$, we must have either $M_1 \succeq_{C_2} M_2$ or $M_2 \succeq_{C_2} M_1$. In the former case, the parents of C_2 in M_1 are the parents of C_2 in $M_1 \oplus M_2$. In particular, if D is the parent of C_2 on the path from C_1 to C_2 in M_1, then D is a parent of C_2 in $M_1 \oplus M_2$. Since none of the intermediate nodes on the path from C_1 to D in M_1 are in M_2 except for C_1, it follows by our earlier argument than the path from C_1 to D in M_1

is also a path from C_1 to D in $M_1 \oplus M_2$. Thus, C_1 is an ancestor of C_2 in $M_1 \oplus M_2$. In the latter case, the parents of C_2 in M_1 must also be in M_2 (in fact, they must be M_1-immediate ancestors of C_2 in M_2). Since none of the intermediate nodes on the path from C_1 to C_2 is in M_2, it must be the case that the path from C_1 to C_2 has length 1, and C_1 is a parent of C_2 in M_1. By $\mathrm{MI1}_{M_2,M_1,C_2}$, there is a path from C_1 to C_2 in M_2 none of whose intermediate nodes is in M_1. Then the same argument given for the case that $M_1 \succeq_{C_2} M_2$ shows that this path in M_2 also exists in $M_1 \oplus M_2$. Thus, C_1 is an ancestor of C_2 in $M_1 \oplus M_2$ in this case as well. The fact that C_2 is ancestor of C_m in $M_1 \oplus M_2$ follows from the induction hypothesis. Thus, C_1 is an ancestor of C_m in $M_1 \oplus M_2$.

For the converse, suppose that A and B are nodes in M_1 and A is an ancestor of B in $M_1 \oplus M_2$. We want to show that A is an ancestor of B in M_1. The argument is similar to that above, but slightly simpler. Again, there is a finite path $A_0, \ldots, A_n$ in the causal graph for $M_1 \oplus M_2$, where $A_0 = A$ and $A_n = B$. If none of the intermediate nodes on the path are in M_2 and at most one of A_0 and A_n is in M_2, then our initial argument shows that this path also exists in M_1.

Now suppose that there are $m > 0$ nodes in M_2 on the path from A to B in $M_1 \oplus M_2$, say $C_1, \ldots, C_m$, in that order. Much like before, we show that (i) C_m is an ancestor of B in M_1, (ii) A is an ancestor of C_1 in M_1, and (iii) C_1 is an ancestor of C_m in M_1. And again, parts (i) and (ii) follow from the earlier argument, since there are no intermediate nodes in M_2 on the path from C_m to B or the path from A to C_1. For part (iii), we again proceed by induction on m. If $m = 1$, the result is trivially true. So suppose that $m > 1$. Since M_1 and M_2 are compatible and C_2 is a node in both M_1 and M_2, again, either $M_1 \succeq_{C_2} M_2$ or $M_2 \succeq_{C_2} M_1$. In the former case, the parents of C_2 in M_1 are just the parents of C_2 in $M_1 \oplus M_2$, so if D is the parent of C_2 on the path from C_1 to C_2 in $M_1 \oplus M_2$, D is a parent of C_2 in M_1. Since the path from C_1 to D in $M_1 \oplus M_2$ has no intermediate nodes in M_2, we can apply earlier argument to show that there is a path from C_1 to D in M_1, and complete the proof as before. In the latter case, all the parents of C_2 in $M_1 \oplus M_2$ must be in M_2, so the path has length 1 and C_1 is a parent of C_2 in $M_1 \oplus M_2$ and in M_2. Thus, C_1 is an immediate M_1-ancestor of C_2 in M_2. $\mathrm{MI1}_{M_2,M_1,C_2}$ implies that C_1 must be a parent of C_2 in M_1. Again, we can complete the proof as before.

The acyclicity of $M_1 \oplus M_2$ is now almost immediate. For suppose that there is a cycle $A_0, \ldots, A_n$ in the causal graph for $M_1 \oplus M_2$, where $A_0 = A_n$ and $n > 0$. Either A_n and A_{n-1} are both in M_1 (if $\mathcal{F}_{1,2}(A_n) = \mathcal{F}_1(A_n)$) or they are both in M_2 (if $\mathcal{F}_{1,2}(A_n) = \mathcal{F}_2(A_n)$). Suppose that they are both in M_1. Then, since A_{n-1} is an ancestor of A_n in $M_1 \oplus M_2$ and A_n is an ancestor of A_{n-1} in $M_1 \oplus M_2$, by the preceding argument, A_{n-1} is an ancestor of A_n in M_1 and A_n is an ancestor of A_{n-1} in M_1, contradicting the acyclicity of M_1. A similar argument applies if both A_{n-1} and A_n are in M_2.

For part (e), suppose that $(M_1, \vec{u}_1)$ and $(M_2, \vec{u}_2)$ are compatible, $\vec{u}$ is a context for $M_1 \oplus M_2$ that agrees with $\vec{u}_1$ on the variables in $\mathcal{U}_1 \cap \mathcal{U}_2$, and φ is a formula that mentions only variables in M_1. It clearly suffices to show that $(M_1, \vec{u}_1) \models \varphi$ iff $(M_1 \oplus M_2, \vec{u}) \models \varphi$ if φ has the form $[\vec{X} \leftarrow \vec{x}](Y = y)$, where $(\vec{X} \cup \{Y\}) \subseteq \mathcal{V}_1$. To show this, it suffices to show that $((M_1)_{\vec{X}=\vec{x}}, \vec{u}_1) \models (Y = y)$ iff $((M_1 \oplus M_2)_{\vec{X}=\vec{x}}, \vec{u}) \models (Y = y)$. Define the *depth* of a variable Y in a causal graph to be the length of the longest path from an exogenous variable to Y in the graph. We prove, by induction on the depth of the variable Y in the causal graph of $M_1 \oplus M_2$, that for all contexts $\vec{u}_1$ in M_1, $\vec{u}_2$ in M_2, and $\vec{u}$ in

$M_1 \oplus M_2$, (i) if $\vec{X} \subseteq \mathcal{U}_1 \cup \mathcal{V}_1$, $Y \in \mathcal{V}_1$, and $\vec{u}$ and $\vec{u}_1$ agree on the variables in $\mathcal{U} \cap \mathcal{U}_1$, then $((M_1)_{\vec{X}=\vec{x}}, \vec{u}_1) \models (Y = y)$ iff $((M_1 \oplus M_2)_{\vec{X}=\vec{x}}, \vec{u}) \models (Y = y)$, and (ii) if $\vec{X} \subseteq \mathcal{U}_2 \cup \mathcal{V}_2$, $Y \in \mathcal{V}_2$, and $\vec{u}$ and $\vec{u}_2$ agree on the variables in $\mathcal{U} \cap \mathcal{U}_2$, then $((M_2)_{\vec{X}=\vec{x}}, \vec{u}_2) \models (Y = y)$ iff $((M_1 \oplus M_2)_{\vec{X}=\vec{x}}, \vec{u}) \models (Y = y)$. (Note that if $Y \in (\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2)$, then it must satisfy both (i) and (ii).)

If Y has depth 0, then Y is an exogenous variable, which is inconsistent with our assumption that Y is an endogenous variable. If Y has depth $d > 0$, we consider a number of cases. First, observe that the result holds trivially if $Y \in \vec{X}$, so we can assume that $Y \notin \vec{X}$. If $Y \in \mathcal{V}_1 - (\mathcal{U}_2 \cup \mathcal{V}_2)$, then the parents of Y in $M_1 \oplus M_2$ are the same as the parents of Y in M_1, so (i) is immediate from the induction hypothesis and (ii) is vacuously true. Similarly, if $Y \in \mathcal{V}_2 - (\mathcal{U}_1 \cup \mathcal{V}_1)$, then (ii) is immediate from the induction hypothesis and (i) is vacuously true. If $Y \in (\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2)$ and $(M_1, \vec{u}_1) \succeq_Y (M_2, \vec{u}_2)$, then again, the parents of Y in $M_1 \oplus M_2$ are the same as the parents of Y in M_1, so (i) is immediate from the induction hypothesis. To show that (ii) holds, fix appropriate contexts $\vec{u}_2$ and $\vec{u}$. Now the parents of Y in M_2 are the immediate M_2-ancestors of Y in M_1. Let $\vec{Z} = Par_{M_2}(Y)$. It follows from the arguments for part (c) that for all $Z \in \vec{Z}$, all the paths from Z to Y in M_1 also exist in $M_1 \oplus M_2$ and the parents of Y in M_2 are exactly the immediate M_2-ancestors of Y in $M_1 \oplus M_2$. That is, $\vec{Z}$ screens Y from all other variables in M_2 not only in M_2, but also in M_1 and $M_1 \oplus M_2$. Suppose that $((M_2)_{\vec{X} \leftarrow x}, \vec{u}_2) \models \vec{Z} = \vec{z}$. It follows from the induction hypothesis that $((M_1 \oplus M_2)_{\vec{X}=\vec{x}}, \vec{u}) \models \vec{Z} = \vec{z}$. Let $\vec{W} = ((\mathcal{U}_1 \cup \mathcal{V}_1) \cap (\mathcal{U}_2 \cup \mathcal{V}_2)) - \{Y\}$. Let $\vec{w}$ be a setting for $\vec{W}$ that agrees with $\vec{z}$ on the variables in $\vec{Z}$. Then we have the following chain of equivalences:

$$
\begin{array}{ll}
 & ((M_2)_{\vec{X}=\vec{x}}, \vec{u}) \models Y = y \\
\text{iff} & ((M_2)_{\vec{X}=\vec{x}}, \vec{u}_2) \models [\vec{Z} \leftarrow \vec{z}](Y = y) \\
\text{iff} & ((M_2)_{\vec{X}=\vec{x}}, \vec{u}_2) \models [\vec{W} \leftarrow \vec{w}](Y = y) \\
\text{iff} & (M_2, \vec{u}_2) \models [\vec{W} \leftarrow \vec{w}](Y = y) \\
\text{iff} & (M_1, \vec{u}_1) \models [\vec{W} \leftarrow \vec{w}](Y = y) \quad [\text{by MI4}_{(M_1,\vec{u}_1),(M_2,\vec{u}_2),Y}] \\
\text{iff} & (M_1, \vec{u}_1) \models [\vec{Z} \leftarrow \vec{z}](Y = y) \\
\text{iff} & ((M_1)_{\vec{Z}=\vec{z}}, \vec{u}_1) \models (Y = y) \\
\text{iff} & ((M_1 \oplus M_2)_{\vec{Z}=\vec{z}}, \vec{u}_1) \models (Y = y) \quad [\text{already shown}] \\
\text{iff} & ((M_1 \oplus M_2), \vec{u}_1) \models [\vec{Z} \leftarrow \vec{z}](Y = y) \\
\text{iff} & ((M_1 \oplus M_2)_{\vec{X}=\vec{x}}, \vec{u}_1) \models [\vec{Z} \leftarrow \vec{z}](Y = y) \\
\text{iff} & ((M_1 \oplus M_2)_{\vec{X}=\vec{x}}, \vec{u}_1) \models Y = y \\
 & \quad [\text{since } (M_1 \oplus M_2)_{\vec{X}=\vec{x}}, \vec{u}_1) \models \vec{Z} = \vec{z}]
\end{array}
$$

The argument is symmetric if $(M_2, \vec{u}_2) \succeq_Y (M_1, \vec{u}_1)$. This completes the proof of (e).

Part (f) is immediate from the definitions.

For part (g), suppose that $M_1 = ((\mathcal{U}_1, \mathcal{V}_1, \mathcal{R}_1), \mathcal{F}_1)$, $M_2 = ((\mathcal{U}_2, \mathcal{V}_2, \mathcal{R}_2), \mathcal{F}_2)$, $M_3 = ((\mathcal{U}_3, \mathcal{V}_3, \mathcal{R}_3), \mathcal{F}_3)$, $M_1 \oplus M_2 = ((\mathcal{U}_{1,2}, \mathcal{V}_{1,2}, \mathcal{R}_{1,2}), \mathcal{F}_{1,2})$, $M_2 \oplus M_3 = ((\mathcal{U}_{2,3}, \mathcal{V}_{2,3}, \mathcal{R}_{2,3}), \mathcal{F}_{2,3})$, $M_1 \oplus (M_2 \oplus M_3) = ((\mathcal{U}_{1,2,3}, \mathcal{V}_{1,2,3}, \mathcal{R}_{1,2,3}), \mathcal{F}_{1,2,3})$, and $(M_1 \oplus M_2) \oplus M_3 = ((\mathcal{U}'_{1,2,3}, \mathcal{V}'_{1,2,3}, \mathcal{R}'_{1,2,3}), \mathcal{F}'_{1,2,3})$. We want to show that $M_1 \oplus (M_2 \oplus M_3) = (M_1 \oplus M_2) \oplus M_3$. It is almost immediate from the definitions that $\mathcal{U}_{1,2,3} = \mathcal{U}'_{1,2,3}$, $\mathcal{V}_{1,2,3} = \mathcal{V}'_{1,2,3}$, and $\mathcal{R}_{1,2,3} = \mathcal{R}'_{1,2,3}$. To show that $\mathcal{F}_{1,2,3} = \mathcal{F}'_{1,2,3}$, we show that for all variables $C \in \mathcal{V}_{1,2,3}$, $\mathcal{F}_{1,2,3}(C) = \mathcal{F}'_{1,2,3}(C)$. We proceed by cases. First suppose that C is in exactly one of the models. For ex-

ample, C is in M_1 but not M_2 or M_3 (i.e., $C = (\mathcal{U}_1 \cup \mathcal{V}_1) - (\mathcal{U}_2 \cup \mathcal{V}_2 \cup \mathcal{U}_3 \cup \mathcal{V}_3)$). If $C \in \mathcal{U}_1$, there is nothing further to prove. If $C \in \mathcal{V}_1$, then it is easy to check that $\mathcal{F}_{1,2,3}(C) = \mathcal{F}'_{1,2,3}(C) = \mathcal{F}_1(C)$. The same argument works if C is just in M_2 or just in M_3.

If C is in two of the three models, suppose without loss of generality that C is in M_1 and M_2 but not M_3. Note that if M and M are compatible, then we must have either $M \succeq_C M'$ or $M' \succeq_C M$; moreover, $M \succ_C M'$ iff $\mathrm{MI1}_{M',M,C}$ does not hold. Going back to the proof, since M_1 and M_2 are compatible, as we observed, either $M_1 \succeq_C M_2$ or $M_2 \succeq_C M_1$ (or both). If $M_1 \succeq_C M_2$ then either $C \in \mathcal{U}_1 \cap \mathcal{U}_2$, in which case there is nothing further to prove, or $C \in \mathcal{V}_1$. In that case, $\mathcal{F}_{1,2}(C) = \mathcal{F}_1(C)$, so $\mathcal{F}_{1,2,3}(C) = \mathcal{F}_1(C)$. Since $C \notin \mathcal{V}_3$, we have $\mathcal{F}_{2,3}(C) = \mathcal{F}_2(C)$. If we also have $M_2 \succeq_C M_1$, then by (a), $\mathcal{F}_1(C) = \mathcal{F}_2(C)$, and it is easy to see that $\mathcal{F}'_{1,2,3}(C) = \mathcal{F}_1(C)$. Now suppose that $M_2 \not\succeq_C M_1$. As we observed, this means that $MI1_{M_2,M_1,C}$ does not hold. M_1 is compatible with $M_2 \oplus M_3$, so we must either $M_1 \succeq_C M_2 \oplus M_3$ or $M_2 \oplus M_3 \succeq_C M_1$. It is easy to see that since $M_2 \not\succeq_C M_1$, we cannot have $M_2 \oplus M_3 \succeq_C M_1$, so we must have $M_1 \succeq_C M_2 \oplus M_3$. It follows that $\mathcal{F}'_{1,2,3}(C) = \mathcal{F}_1(C)$.

Finally, suppose that C is in all three models. We first show that $\succeq_C$ is transitive when restricted to M_1, M_2, and M_3. For suppose that $M_1 \succeq_C M_2$ and $M_2 \succeq_C M_3$. If $M_1 \sim_C M_2$ or $M_2 \sim_C M_3$, then it is easy to see that $M_1 \succeq_C M_3$. So suppose that $M_1 \succ_C M_2$ and $M_2 \succ_C M_3$. Since M_1 and M_3 are compatible, we must have either $M_1 \succeq_C M_3$ or $M_3 \succeq_C M_1$.

Suppose by way of contradiction that $M_3 \succ_C M_1$. Let $\vec{X}_1 = Par_{M_1}(C)$, $\vec{X}_2 = Par_{M_2}(C)$, and $\vec{X}_3 = Par_{M_3}(C)$. We now construct an infinite sequence of variables $A_0, A_1, \ldots$ such that each variable in the sequence is either in $\vec{X}_2 - \vec{X}_1$, $\vec{X}_3 - \vec{X}_2$, or $\vec{X}_1 - \vec{X}_3$, and if variable A_n is in $\vec{X}_i - \vec{X}_j$, then the next variable is in $\vec{X}_j$ and there is a path in M_j from A_n to A_{n+1}. We proceed by induction. Since $M_1 \succ_C M_2$, by $\mathrm{MI1}_{M_1,M_2,C}$ there must be at least one variable in $A_0 \in \vec{X}_2 - \vec{X}_1$ and a path from Z_1 to C in M_1 that does not go through any other variables in $\vec{X}_2$. Since $\vec{X}_1$ screens C from all ancestors in M_1, this path must go through a variable $A_1 \in \vec{X}_1 - \vec{X}_2$. If $A_1 \in \vec{X}_3$, then it is in $\vec{X}_3 - \vec{X}_2$; if $A_1 \notin \vec{X}_3$, it is in $\vec{X}_1 - \vec{X}_3$. Either way, A_1 is an appropriate successor of A_0 in the sequence. The inductive step of the argument is identical; if $A_n \in \vec{X}_i - \vec{X}_j$, we use the fact that $M_j \succ_C M_i$ to construct A_{n+1}. Note that, for all $n \geq 0$, since $A_n \in \vec{X}_i - \vec{X}_j$ and $A_{n+1} \in \vec{X}_j$, we must have $A_n \neq A_{n+1}$. Moreover, by the argument in the proof of (c) since there is a path from A_n to A_{n+1} in M_j, there must also be such a path in $M_1 \oplus (M_2 \oplus M_3)$. Since there are only finitely many variables altogether, there must be some N_1 and N_2 such that $A_{N_1} = A_{N_2}$. That means we have a cycle in $M_1 \oplus (M_2 \oplus M_3)$, contradicting (c).

Since $\succeq_C$ is transitive and complete on $\{M_1, M_2, M_3\}$ (completeness says that for each pair, one of the two must be dominant), one of M_1, M_2, and M_3 must dominate the other two with respect to $\succeq_C$. Suppose it is M_1. It is easy to see that $M_1 \oplus M_2 \succeq_C M_3$ and $M_1 \succeq_C (M_2 \oplus M_3)$. It then easily follows that $\mathcal{F}_{1,2,3}(C) = \mathcal{F}'_{1,2,3}(C) = \mathcal{F}_1(C)$. A similar argument holds if M_2 or M_3 is the model that dominates with respect to $\succeq_C$. ∎

Acknowledgments: We thank Noemie Bouhana, Frederick Eberhardt, Meir Friedenberg, and anonymous reviewers for useful comments. Joe Halpern's work was supported by NSF grants IIS-1703846 and IIS-1718108, AFOSR grant FA9550-12-1-0040, ARO

grant W911NF-17-1-0592, and the Open Philanthropy project. Dalal Alrajeh's work was supported by MRI grant FA9550-16-1-0516.

References

[1] Bradley, R., Dietrich, F., List, C., 2014. Aggregating causal judgments. Philosophy of Science 81, 419–515.

[2] Chockler, H., Fenton, N.E., Keppens, J., Lagnado, D.A., 2015. Causal analysis for attributing responsibility in legal cases, in: Proc. 15th International Conference on Artificial Intelligence and Law (ICAIL '15), pp. 33–42.

[3] Claassen, T., Heskes, T., 2010. Learning causal network structure from multiple (in)dependence models, in: Proc. of the Fifth European Workshop on Probabilistic Graphical Models, pp. 81–88.

[4] Claassen, T., Heskes, T., 2012. A Bayesian approach to constraint based causal inference, in: Proc. 28th Conference on Uncertainty in Artificial Intelligence (UAI 2012), pp. 207–217.

[5] Dawid, A., 1987. The difficulty about conjunction. Journal of the Royal Statistical Society, Series D 36, 91–97.

[6] Feng, G., Zhang, J., Liao, S.S., 2014. A novel method for combining Bayesian networks, theoretical analysis, and its applications. Pattern Recognition 47, 2057–2069.

[7] Fenton, N., Neil, M., Berger, D., 2016. Bayes and the law. Annual Review of Statistics and Its Application 3, 51–77.

[8] Friedenberg, M., Halpern, J.Y., 2018. Combining the causal judgments of experts with possibly different focus areas, in: Principles of Knowledge Representation and Reasoning: Proc. Sixteenth International Conference (KR '18).

[9] Glymour, C., Wimberly, F., 2007. Actual causes and thought experiments, in: Campbell, J., O'Rourke, M., Silverstein, H. (Eds.), Causation and Explanation. MIT Press, Cambridge, MA, pp. 43–67.

[10] Halpern, J.Y., 2015. A modification of the Halpern-Pearl definition of causality, in: Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), pp. 3022–3033.

[11] Halpern, J.Y., 2016a. Actual Causality. MIT Press, Cambridge, MA.

[12] Halpern, J.Y., 2016b. Appropriate causal models and stability of causation. Review of Symbolic Logic 9, 76–102.

[13] Halpern, J.Y., Pearl, J., 2005. Causes and explanations: a structural-model approach. Part I: Causes. British Journal for Philosophy of Science 56, 843–887.

[14] Hitchcock, C., 2001. The intransitivity of causation revealed in equations and graphs. Journal of Philosophy XCVIII, 273–299.

[15] Hitchcock, C., 2007. Prevention, preemption, and the principle of sufficient reason. Philosophical Review 116, 495–532.

[16] Hoover, K.D., 2008. Causality in economics and econometrics, in: Blume, L., Durlauf, S. (Eds.), The New Palgrave: A Dictionary of Economics. Palgrave Macmillan, New York.

[17] Hyttinen, A., Eberhardt, F., Jarvisalo, M., 2014. Constraint-based causal discovery: conflict resolution with answer set programming, in: Proceedings of the Thirtieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-14), AUAI Press, Corvallis, Oregon. pp. 340–349.

[18] Illari, P.M., Russo, F., Williamson, J. (Eds.), 2011. Causality in the Sciences. Oxford University Press, Oxford, U.K.

[19] Jenner, B., Toran, J., 1995. Computing functions with parallel queries to NP. Theoretical Computer Science 141, 175–193.

[20] Johnson, D.S., 1990. A catalog of complexity classes, in: Leeuwen, J.v. (Ed.), Handbook of Theoretical Computer Science. Elsevier Science. volume A. chapter 2.

[21] Lewis, D., 2000. Causation as influence. Journal of Philosophy XCVII, 182–197.

[22] Pearl, J., 1993. Comment: Graphical models, causality and intervention. Statistical Science 8, 266–269.

[23] Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, New York.

[24] Spirtes, P., Glymour, C., Scheines, R., 1993. Causation, Prediction, and Search. Springer-Verlag, New York.

[25] Tillman, R.E., Spirtes, P., 2011. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables, in: Proceedings of the Fourteenth

International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011, JMLR.org. pp. 3–15.
[26] Triantafillou, S., Tsamardinos, I., 2015. Constraint-based causal discovery from multiple interventions over overlapping variable sets. Journal of Machine Learning Research 16, 2147–2205.
[27] Wagner, K.W., 1990. Bounded query classes. SIAM J. Comput. 19, 833–846.
[28] Woodward, J., 2003. Making Things Happen: A Theory of Causal Explanation. Oxford University Press, Oxford, U.K.
[29] Zennaro, F.M., Ivanovska, M., 2018a. Counterfactually fair prediction using multiple causal models, in: Multi-Agent Systems - 16th European Conference, EUMAS 2018, Revised Selected Papers, Springer. pp. 249–266.
[30] Zennaro, F.M., Ivanovska, M., 2018b. Pooling of causal models under counterfactual fairness via causal judgement aggregation. Www.arxiv.org.