Combining Experts' Causal Judgments

Dalal Alrajeh
Department of Computing
Imperial College London
dalal.alrajeh04@imperial.ac.uk

Hana Chockler
Department of Informatics
King's College London
Hana.Chockler@kcl.ac.uk

Joseph Y. Halpern
Computer Science Department
Cornell University
halpern@cs.cornell.edu

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 combined. We then use it as the basis for combining experts causal judgments. 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. (Genest and Zidek (1986) provide a somewhat dated but still useful overview; Dawid (1987) and Fenton et al. (2016), 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 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., (Chockler et al. 2015; Sampson, Winship, and Knight 2013)), there is surprisingly little work on combining causal judgments. Indeed, the only work that we are aware of is that of Bradley, Dietrich, and List (2014) (BDL from now on), who prove an impossibility result. Specifically, they describe certain arguably reasonable desiderata, and show that there is no way of combining 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, none of which seem that satisfactory.

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., (Claassen and Heskes 2010; 2012; Hyttinen, Eberhardt, and Jarvisalo 2014; Tillman and Spirtes 2011; Triantafillou and Tsamardinos 2015); Triantafillou and Tsamardinos (2015) 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 causals model 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 combining experts' causal models when sufficient data for discovering the overall causal model is not available. The key step in combining experts' causal models lies in defining when two causal models are *compatible*. Causal models can be combined 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,

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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 combined 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).

This set of constraints is very restrictive, and, as we show on real-life examples, models are often not compatible in this strong sense. We thus define two successively more general notions of compatibility. But even with this more general notions, we may find that not all the experts' models are incompatible. In that case, we simply place a probability on possible ways of combining the compatible models, using relatively standard techniques, based on the perceived reliability of the experts who proposed them. 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, she 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 believe that our approach provides a useful formal framework that can be applied to the determination of appropriate interventions in real-world scenarios involving complex sociological phenomena, such as crime prevention scenarios (Sampson, Winship, and Knight 2013) and radicalization (Wikström and Bouhana 2017). Proofs and detailed descriptions of some of the examples in the paper are deferred to the full paper, due to lack of space.

2 Causal Models

In this section, we review the definition of causal models introduced by Halpern and Pearl (2005). The material in this section is largely taken from (Halpern 2016).

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. The structural equations describe how these values are determined.

Formally, a *causal model* M is a pair $(\S, \mathcal{F})$, where $\S$ 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 $\S$ 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 $\mathcal{U} \cup \mathcal{V}$.

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 $M = (\S, \mathcal{F})$ results in a new causal model, denoted $M_{X \leftarrow x}$, which is identical to M, except that the equation 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. There is a directed edge from variable 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 $F_Y(x, \vec{z}) \neq F_Y(x', \vec{z})$.[1] 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). In this paper, following the literature, we restrict to recursive models.

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 $\S$)* 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 $\mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$. 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 $\vec{u}$ (that is, the unique vector of values for the exogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y} = \vec{y}}, \vec{u}) \models \varphi$.

[1] In many papers (e.g., (Bradley, Dietrich, and List 2014; Sampson, Winship, and Knight 2013)), a causal model is defined by a causal graph indicating the dependencies, perhaps with an indication of whether a change has a positive or negative effect. Our models are more expressive, since the equations typically provide much more detailed information regarding the dependency between variables.

3 Interventions

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

Definition 1 $\vec{X} = \vec{x}$ *is an* intervention on φ 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 slightly generalizes others in the literature. Pearl (2009) 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. Lu and Druzdzel (2002), Korb et al. (2004), and Woodward (2003) take similar approaches.

We do not require special intervention variables; we just allow interventions directly on the variables in the model. But we can 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. It should be clear that all these definitions are trying to capture the same intuitions, and differ only in minor ways.

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 (Glymour and Wimberly 2007; Halpern 2015; Halpern and Pearl 2005; Hitchcock 2001; 2007; Woodward 2003), there are a number of subtle variations. Fortunately, 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 causality given by Halpern (2015). It has conditions AC1–3 that are analogous 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. The following example, due to Lewis (2000), illustrates some of the subtleties, and highlights the differences between causality and intervention.

Suppose that 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 Suzy not thrown. Most people would say that Suzy is a cause, and not Billy. Part of the difficulty in getting a good definition of causality is to ensure that the definition gives us this result (given an appropriate causal model). However, Suzy's throw by itself is not an intervention for the bottle shattering. Even if we prevent Suzy from throwing, the bottle will still shatter because of Billy's throw. That is, if we have variables *ST* and *BT* for Suzy's throw and Billy's throw, with possible values 0 and 1 (*ST* $= 1$ if Suzy throws, *ST* $= 0$ if she doesn't, and similarly for Billy), then although *ST* $= 1$ is a cause of the bottle shattering, *ST* $= 0$ is not an intervention for the bottle shattering; intervening on *ST* alone does not change the outcome. On the other hand, *ST* $= 0 \wedge$ *BT* $= 0$ is an intervention for the bottle shattering, but *ST* $= 1 \wedge$ *BT* $= 1$ is not a cause of the bottle shattering; it violates AC3.

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

Proposition 3.1 *If* $\vec{X} = \vec{x}$ *is an intervention for* φ *in* $(M, \vec{u})$ *then there is some subset of* $\vec{X}'$ *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 for* φ*.*

Halpern (2015) 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 $L = L_1 \cap L_2$ (Papadimitriou and Yannakakis 1982). It is well known that *DP* is at least as hard as NP and co-NP (and most likely strictly harder). The problem of determining whether $\vec{X} = \vec{x}$ is an intervention is in a lower complexity class.

Theorem 3.2 *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 for* φ *in* $(M, \vec{u})$ *is co-NP-complete.*

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 $O(|M|^k)$.

4 Combining Compatible Causal Models

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 5.) We start with a strong notion of compatibility, and then consider generalizations of this notion that are more widely applicable.

4.1 Domination and 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 as far as the variables that M_1 and M_2 share, but M_1 contains (possibly) more detailed information about C, because, for example, there are additional variables in M_1 that affect C. We capture this property formally below. 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 2 *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 (the causal graph corresponding to) 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 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,*

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

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 MI2$_{M_1,M_2,C}$ and MI3$_{M_1,M_2,C}$. 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 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 observations: First, note that there is an abuse of notation in the statement of MI3$_{M_1,M_2,C}$. We allow the set $\vec{X}$ in the statement of MI3$_{M_1,M_2,C}$ to include exogenous variables. However, in giving the semantics of the causal language, we consider only formulas of the form $[\vec{X} \leftarrow \vec{x}]\varphi$ where $\vec{X}$ mentions only endogenous variables. (Note that it is possible that some variables that are exogenous in M_1 may be endogenous in M_2, and vice versa.) Suppose that $\vec{X} \cap \mathcal{U}_1 = \mathcal{U}_1'$ and $\vec{X}' = \vec{X} - \mathcal{U}_1'$; then by $(M_1, \vec{u}_1) \models [\vec{X} \leftarrow \vec{x}](C = c)$ we mean $(M_1, \vec{u}_1') \models [\vec{X}' \leftarrow \vec{x}'](C = c)$, where $\vec{x}'$ is $\vec{x}$ restricted to the variables in $\vec{X}'$, and $\vec{u}_1'$ agrees with $\vec{u}_1$ on the variables in $\mathcal{U}_1 - \mathcal{U}_1'$, and agrees with $\vec{x}$ on the variables in $\mathcal{U}_1'$. 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 4.1), which is the only context in which we are interested in transitivity.

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 *strongly 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 combine M_1 and M_2 by taking the equations for C to be determined by the model that has more information about C.

Example 1 *(Bradley, Dietrich, and List 2014) 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 1, 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. In these graphs (as in many other causal graphs in the literature), exogenous variables are omitted; all variables are taken to be endogenous. Neither MI1$_{M_1,M_2,F}$ nor MI1$_{M_2,M_1,F}$ holds,*

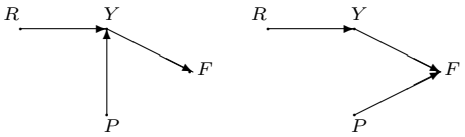

Figure 1: Two expert models of famine.

since P is not an M_2-immediate ancestor of F in M_1. Similarly, neither MI1$_{M_1,M_2,Y}$ nor 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). 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); 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 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 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 we may want to com-

bine models. Consider the following example, taken from the work of Sampson, Winship, and Knight (2013).

Example 2 *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). Sampson et al. (2013) use the following causal graphs shown in Figure 2, which are annotated with the direction of the influence (the only information provided by the experts) to describe the expert's opinions.

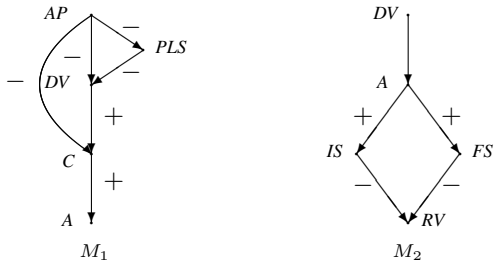

Figure 2: Expert's models of domestic violence.

For the two common variables DV and A, MI1$_{M_1,M_2,DV}$ and 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 MI2$_{M_1,M_2,DV}$ holds. Sampson et al. seem to be implicitly assuming this, and that MI3 holds, so they combine M_1 and M_2 to get the causal graph shown in Figure 3.

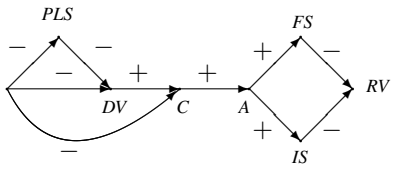

Figure 3: Combined experts' model of domestic violence.

Note that model M_2 in Figure 2 does not state how DV influences A. Presumably, this respresents the expert's undertainty. 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 5 how such uncertainty can be handled. ∎

Suppose that some parent of *AP* (or *AP* itself) in M_1 is not a parent of *DV* in M_2. Then, in M_1, it may be possible to change the value of *DV* by intervening on *AP*, while keeping the values of all the exogenous variables that are parents of *DV* in M_2 fixed. This will seem like an inexplicable change in the value of *DV* from the perspective of the second expert. If the second expert had been aware of such possible changes, she surely would have added additional variables to M_2 to capture this situation. One explanation of the fact that no changes were observed is that the second expert was working in a setting where the values of all variables that she cannot affect by an intervention are determined by some default setting of exogenous variables of which she is not aware (or not modeling). We now define a notion of domination that captures this intuition.

Definition 3 *Let $\vec{v}^*$ be a default setting for the variables in M_1 and M_2. M_1* weakly dominates *M_2 with respect to C relative to $\vec{v}^*$, denoted $M_1 \succeq_C^{\vec{v}^*} M_2$, if MI1$_{M_1,M_2,C}$ holds, 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$_{M_1,M_2,C,\vec{v}^*}$ *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 $\vec{X}$, all values c of C, and all contexts $\vec{u}_1$ for M_1 and $\vec{u}_2$ for M_2 such that $\vec{u}_1$ and $\vec{u}_2$ agree on the variables in $\mathcal{U}_1 \cap \mathcal{U}_2$, $\vec{u}_1$ agrees with $\vec{v}^*$ on the variables in $\mathcal{U}_1 - \mathcal{U}_2$, and $\vec{u}_2$ agrees with $\vec{v}^*$ on the variables in $\mathcal{U}_2 - \mathcal{U}_1$, we have*

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

It is easy to see that $\succeq_C$ is a special case of $\succeq_C^{\vec{v}^*}$:

Lemma 1 *If $M_1 \succeq_C M_2$, then, for all default settings $\vec{v}^*$ of the variables in M_1 and M_2, we have $M_1 \succeq_C^{\vec{v}^*} M_2$.*

In light of Lemma 1, we give all the definitions in the remainder of the paper using $\succeq_C^{\vec{v}^*}$. All the technical results hold if we replace $\succeq_C^{\vec{v}^*}$ by $\succeq_C$ throughout.

Definition 4 *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) 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 (2) for all variables $C \in \mathcal{V}_1 \cap \mathcal{V}_2$, either $M_1 \succeq_C^{\vec{v}^*} M_2$ or $M_2 \succeq_C^{\vec{v}^*} M_1$. If M_1 and M_2 are compatible, then $M_1 \oplus M_2$ 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 $M_1 \succeq_C^{\vec{v}^*} M_2$, 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 $M_2 \succeq_C^{\vec{v}^*} M_1$, then $\mathcal{F}(C) = \mathcal{F}_2(C)$.*[2]

Returning to Example 2, assume that either MI2$_{M_1,M_2,DV}$, MI2$_{M_1,M_2,A}$, MI3$_{M_1,M_2,DV}$, and MI3$_{M_1,M_2,A}$ all hold, or there is a default setting $\vec{v}^*$

[2]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.

such that $\text{MI4}_{M_1,M_2,DV,\vec{v}^*}$ and $\text{MI4}_{M_1,M_2,A,\vec{v}^*}$ hold. Then $M_1 \oplus M_2$ has the causal graph described in Figure 3; that is, even though Sampson et al. (2013) do not have a formal theory for combining models, they actually combine models in just the way that we are suggesting.

Let $M_1 \sim_C^{\vec{v}^*} M_2$ be an abbreviation for $M_1 \succeq_C^{\vec{v}^*} M_2$ and $M_2 \succeq_C^{\vec{v}^*} M_1$. We also write $M_1 \succ_C^{\vec{v}^*} M_2$ if $M_1 \succeq_C^{\vec{v}^*} M_2$ and $M_2 \not\succeq_C^{\vec{v}^*} M_1$.

The next proposition provides evidence that Definition 4 is reasonable and captures our intuitions. Part (b) says that it 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 $M_1 \sim_C^{\vec{v}^*} M_2$, there is in fact no conflict; part (a) is a technical result needed to prove part (b). Part (c) states that the combined model in the compatible case 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, provided that the default values are used for the exogenous variables not in $\mathcal{U}_1 \cap \mathcal{U}_2$. Parts (d) and (e) can be viewed as saying that the essential causal structure of M_1 is preserved in $M_1 \oplus M_2$. Finally, parts (f) and (g) say that $\oplus$ is commutative and associative over its domain.

Proposition 4.1 *Suppose that M_1 M_2, and M_3 are pairwise compatible. Then the following conditions hold.*

(a) If $M_1 \sim_C^{\vec{v}^*} M_2$ 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.

(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 $M_1 \oplus M_2$.

(e) If φ is a formula that mentions only the endogenous variables in M_1, $\vec{u}$ is a context for $M_1 \oplus M_2$, $\vec{u}_1$ is a context for M_1, $\vec{u}$ and $\vec{u}_1$ agree on the variables in $\mathcal{U}_1 \cap \mathcal{U}_2$, $\vec{u}$ agrees with $\vec{v}^*$ on the variables in $\mathcal{U} - (\mathcal{U}_1 \cap \mathcal{U}_2)$, and $\vec{u}_1$ agrees with $\vec{v}^*$ on the variables in $\mathcal{U}_1 - \mathcal{U}_2$, 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 $M_1 \oplus (M_2 \oplus M_3) = (M_1 \oplus M_2) \oplus M_3$.

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 trivially holds. If $|\mathcal{M}| = 2$, then the models in $\mathcal{M}$ are mutually compatible if they are compatible according to Definition 4. 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 Proposition 4.1, 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. Proposition 4.1(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 combining mutually compatible models, we are maximizing our use of information.

We now discuss the extent to which our approach to combining 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 and can output any logically possible model. This clearly holds for us.
- *Acyclicity*: the result of combining M_1 and M_2 is acyclic. This follows from Proposition 4.1(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, and This is trivial 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 combined.) 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, so we do not satisfy neutrality.
- *Non-Dictatorship:* no single expert determines the aggregation. This clearly holds for us.

4.2 Partial Compatibility

While the notion of dominance used in Definition 4 is useful, it still does not cover many cases of interest. We briefly describe an example here on causal models for the emergence of radicalization in US prisons. The material is taken from Useem and Clayton (2009). Although Useem and Clayton do not provide causal models, we construct these based on the description provided. We do not provide a detailed explanation of all the variables and their dependencies here (details are provided in the full paper); for our purposes, it suffices to focus on the structure of these models.

Example 3 *Consider the two causal models in Figure 4. The SCICH model represents expert 1's opinion about emerging radicalization (R) in the State Correctional Institution Camp Hill in Pennsylvania. The TDCR model represents expert 2's opinion about the causes of emergence in the Texas Department of Corrections and Rehabilitation (TDCR). Both experts agree on the structural equations for R. However they differ on the structural equations for PD, CB and AM. For the common variables CB and AM, assuming default values for CG, LM and FM in SCICH and for T, D and RG in TDCR, we can show that SCICH $\succeq_{CB}^{\vec{v}^*}$ TDCR and SCICH $\succeq_{AM}^{\vec{v}^*}$ TDCR. However, neither model dominates the other with respect to PD; neither $\text{MI1}_{SCICH,TDCR,PD}$ nor $\text{MI1}_{TDCR,SCICH,PD}$ holds. Therefore the models are incompatible according to Definition 4.*

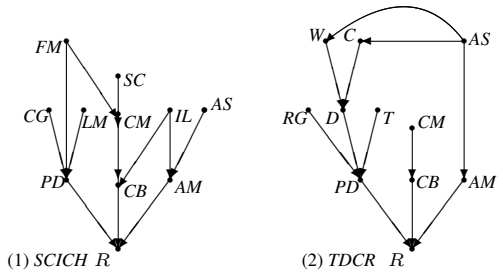

Figure 4: Schematic representation of the two prison models.

Although the models are incompatible according to our definition, the incompatibility is "localized" to the variable *PD*. Moreover, it is not even clear that there is disagreement with regard to *PD*; the experts could just be focusing on different variables. In a richer model, *PD* might have six parents. The trouble is, we have no idea from the two models what the equations for *PD* would be in the richer model.

In the full paper, we define an approach to combining models even when they are not completely compatible. We consider a notion of *partial compatibility* of causal models and construct a partial causal model. Roughly speaking, the causal model does not completely define the equations for variables X that are common to two models, where neither dominates the other with respect to X. In the example above, *PD* has six parents in the partial model, but we do not completely specify the equation for the value of *PD* as a function of its parents' values.

This approach to aggregating models is our main contribution. Using it, we show in the next section how experts' models can be combined to reason about interventions.

5 Combining Experts' 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 degree of confidence that expert i is correct. Let $Compat = \{I \subseteq \{1, \ldots, n\} : \text{the models in } \{M_i : i \in I\} \text{ are mutually compatible}\}$. For $I \in Compat$, define $M_I = \oplus_{i \in I} M_i$. By Proposition 4.1, M_I is well defined. The policymaker considers the models in $\mathcal{M}_{Compat} = \{M_I : I \in Compat\}$, placing the probability of $p_I = \prod_{i \in I}(p_i) * \prod_{j \notin I}(1 - p_j)/N$ on M_I, where $N = \sum_{I \in Compat} p_I$ is a normalization factor.

Intuitively, we view the events "expert i is right" as being mutually independent, for $i = 1, \ldots, n$. Thus, p_I is the probability of the event that exactly the experts in I are right (and the ones not in I are wrong). If exactly the experts in I are indeed right, it seems reasonable to view M_I as the "right" causal model. Note that it is not possible for all the experts in I to be right if there are experts $i, j \in I$ such that M_i and M_j are incompatible. Thus, we consider only models M_I for $I \in Compat$. But even if $I \in Compat$, it is possible that some of the experts in I are wrong in their causal judgments. 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 $I \in Compat$; we place positive probability on $M_{I'}$ even if I' is a strict subset of some $I \in Compat$. This method of combining experts' judgments is similar in spirit to the method proposed by Dawid (1987) and Fenton et al. (2016).

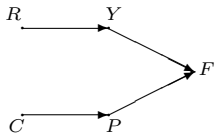

Figure 5: Third expert's (and combined) model of famine.

This completes our description of how to combine experts' causal judgments. At a high level, for each subset of experts whose judgments are compatible (in that the models they are proposing are pairwise compatible), we combine the models, and assign the combined model a probability corresponding the probability of the experts in the subset. Of course, once we have a probability on the settings in $\mathcal{M}_{Compat}$, we can compute, for each setting, which interventions affect the outcome φ of interest, and then compute the probability that a particular intervention is effective.

The straightforward strategy for a policymaker to compute the most effective intervention based on the experts' opinions and the degree of confidence of the policymaker in each expert's judgment is to compute the set $\mathcal{M}_{Compat}$ of models and then to apply the computation of interventions as described in Section 3 to each $M_I \in \mathcal{M}_{Compat}$. The probability that an intervention is effective is computed by summing the probability of the models where it is effective.

To get a sense of how this works, consider a variant of Example 1, 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 5); call this model M_3. According to the compatibility definition in Section 4, the models M_2 and M_3 are compatible (assuming that MI3 holds), but M_1 and M_3 are not. We have $\mathcal{M}_{Compat} = \{\{M_1\}, \{M_2\}, \{M_3\}, \{M_{2,3}\}\}$ 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 and M_1 being wrong. So we have $p_{2,3} = (0.6 * 0.5 * 0.6)/0.56 = 0.32$ (where 0.56 is the normalization factor). In a similar way we compute $p_1 = (0.4 * 0.4 * 0.5)/0.56 = 0.14$. Note that the number of models in $\mathcal{M}_{Compat}$ may be exponential in the number of experts. For example, if all experts are compatible, $Compat$ 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 combined 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.

6 Conclusions

We have provided a method for combining causal models and used that as a basis for combining experts' causal judgments in a way that gets around the impossibility result of Bradley, Dietrich, and List (2014). Our approach can be viewed as a formalization of an earlier work (Chockler et al. 2015; Sampson, Winship, and Knight 2013). 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 for finding the intervention(s) best ameliorate a situation.

Acknowledgments: We thank Noemie Bouhana, Frederick Eberhardt, 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

Bradley, R.; Dietrich, F.; and List, C. 2014. Aggregating causal judgments. *Philosophy of Science* 81(4):419–515.

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

Claassen, T., and Heskes, T. 2010. Learning causal network structure from multiple (in)dependence models. In *Proc. 5th European Workshop on Probabilistic Graphical Models*, 81–88.

Claassen, T., and Heskes, T. 2012. A Bayesian approach to constraint based causal inference. In *Proc. 28th Conf. on Uncertainty in Artificial Intelligence (UAI 2012)*, 207–217.

Dawid, A. 1987. The difficulty about conjunction. *Journal of the Royal Statistical Society, Series D* 36:9197.

Fenton, N.; Neil, M.; and Berger, D. 2016. Bayes and the law. *Annual Review of Statistics and Its Application* 3:51–77.

Genest, C., and Zidek, J. V. 1986. Combining probability distributions: a critique and an annotated bibliography. *Statistical Science* 1(1):114–148.

Glymour, C., and Wimberly, F. 2007. Actual causes and thought experiments. In Campbell, J.; O'Rourke, M.; and Silverstein, H., eds., *Causation and Explanation*. MIT Press. 43–67.

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

Halpern, J. Y. 2015. A modification of the Halpern-Pearl definition of causality. In *Proc. 24th International Joint Conf. on Artificial Intelligence (IJCAI 2015)*, 3022–3033.

Halpern, J. Y. 2016. *Actual Causality*. MIT Press.

Hitchcock, C. 2001. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy* XCVIII(6):273–299.

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

Hyttinen, A.; Eberhardt, F.; and Jarvisalo, M. 2014. Constraint-based causal discovery: conflict resolution with answer set programming. In *Proc. 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, 340–349.

Korb, K. B.; Hope, L. R.; Nicholson, A. E.; and Axnick, K. 2004. Varieties of causal intervention. In *Proc. 8th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence (PRICAI-04)*. 322–331.

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

Lu, T.-C., and Druzdzel, M. J. 2002. Causal models, value of intervention, and search for opportunities. *Advances in Bayesian Networks: Studies in Fuzziness and Soft Computing* 146(30):121–135.

Papadimitriou, C. H., and Yannakakis, M. 1982. The complexity of facets (and some facets of complexity). *Journal of Computer and System Sciences* 28(2):244–259.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition.

Sampson, R. J.; Winship, C.; and Knight, C. 2013. Translating causal claims: Principles and strategies for policy-relevant criminology. *Criminology, Causality, and Public Policy* 12(4):587–616.

Tillman, R. E., and Spirtes, P. 2011. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proc. 14th International Conf. on Artificial Intelligence and Statistics (AISTATS 2011)*, pp. 3–15.

Triantafillou, S., and Tsamardinos, I. 2015. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research* 16:2147–2205.

Useem, B., and Clayton, O. 2009. Radicalization of U.S. prisoners. *Criminology & Public Policy* 561–592.

Wikström, P.-O., and Bouhana, N. 2017. Analysing terrorism and radicalization: A situational action theory. In LaFree, G., and Freilich, J., eds., *The Encyclopaedia of the Criminology of Terrorism*. John Wiley and Sons.

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