The Problem of Granularity for Scientific Explanation

David Kinney

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Department of Philosophy, Logic and Scientific Method

The Problem of Granularity for Scientific Explanation

David Kinney

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics and Political Science for the degree of Doctor of Philosophy, May 2019.

Abstract

This dissertation aims to determine the optimal level of granularity for the variables used in probabilistic causal models. These causal models are useful for generating explanations in a number of scientific contexts. In Chapter 1, I argue that there is rarely a unique level of granularity at which a given phenomenon can be causally explained, thereby rejecting various causal exclusion arguments. In Chapter 2, I consider several recent proposals for measuring the explanatory power of causal explanations, and show that these measures fail to track the comparative depth of explanations given at different levels of granularity. In Chapter 3, I offer a pragmatic account of how to partition the measure space of a causal variable so as to optimally explain its effect. My account uses the decision-theoretic notion of value of information, and indexes the relative depth of an explanation to a particular agent faced with a particular decision problem. Chapter 4 applies this same decision-theoretic framework to answer the epistemic question of how to discover constitutive relationships in nature. In Chapter 5, I describe the formal details of the relationship between random variables that are meant to be coarse-grained and fine-grained representations of the same type of phenomenon. I use this formal framework to rebut a popular argument for the view that special science probabilities can be objective chances. Chapter 6 discusses challenges related to the causal interpretation of Bayes nets that use imprecise rather than precise probabilities.

Declaration

I certify that the thesis I have presented for examination for the degree of Doctor of Philosophy of the London School of Economics and Political Science is solely my own work. I confirm that Chapter 3 is published in *Philosophy of Science* as Kinney (2019), and that Chapter 6 is published in *Information* as Kinney (2018). Small amounts of material from both of these published articles are also included in Chapter 1. This dissertation contains 67,103 words.

David Kinney

Acknowledgements

I should begin by acknowledging the privileges that I have benefited from throughout my life. Over the course of my childhood, my parents became very affluent; because they could afford a home in a leafy suburb of New York City, I was able to attend excellent publicly-funded schools while growing up in comfort and safety. My parents paid my entire undergraduate tuition at Dartmouth College, as well as a portion of my tuition for my MSc degree at LSE. I have also benefited from many of the advantages that come with being a white, heterosexual, cis-gender man who is a native English speaker. While I am proud of the hard work that I put into this dissertation, my path to this point has been drastically eased by these privileges, which I did nothing to earn and over which I had no control.

Since I began my MSc in 2013, Katie Steele has been an extremely supportive advisor and mentor. She has read and provided insightful comments on every word of this dissertation, and my conversations with her have been invaluable with respect to developing ideas for chapters. When Katie left LSE for the Australian National University in 2016, she arranged for me to visit ANU on two separate occasions. These research visits were enormously productive with respect to getting work done on this dissertation. Throughout my PhD, I have always felt as though Katie was in my corner. More than anyone else, I owe my early successes in academia to her.

After Katie left for ANU, Luc Bovens and Jonathan Birch both stepped up to provide additional PhD supervision. Like Katie, they provided insightful comments on drafts of nearly every chapter in this dissertation. These chapters are in much better shape as a result of their hard work. Over a recent spring break, Luc powered through several final drafts of chapters, offering very helpful feedback. As I write this now, Jonathan is currently in the process of giving the entire dissertation a final look over. In addition, Jonathan was an extremely helpful advisor during my time on the job market, which overlapped with some of the time spent completing this dissertation. Jonathan's careful feedback on all of my dossier materials helped me navigate the job market efficiently, which had the knock-on effect of freeing up time and energy to complete the dissertation.

I thank all of the administrative teams at the philosophy departments at LSE and ANU, most notably Andrea Pawley for all her help with job market support. I also thank LSE for financial support over the course of this PhD.

I thank two of my undergraduate professors who kick-started my interest in philosophy: Christine Thomas and Samuel Levey.

Apart from my advisors, conversations with and support from a large number of philosophers and other academics at LSE, ANU, and elsewhere have helped me to improve both the quality of this dissertation and my ability as a philosopher. Many of these people are also my good friends. In alphabetical order, they include: Jason Alexander, Richard Bradley, Liam Kofi Bright, Heather Browning, Rachael Brown, Kamilla Buchter, Susanne Burri, Mark Canciani, Chloe de Canson, Lorenzo Casini, Camilla Colombo, Nicholas Cote, Paul Daniell, Peter Dennis, Hugh Desmond, Chris Dorst, Keith Dowding, Phil Dowe, Christina Easton, Melissa Ebbers, Frederick Eberhardt, Marcus Eronen, Benjamin Eva, Goreti Faria, Alison Fernandes, Fiona Fidler, Roman Frigg, Paul Griffiths, Alan Hájek, Margherita Harris, Todd Karhu, Anton Killin, Kuan Ko-Hung, David Lagnado, Chad Lee-Stronach, Erick Llamas, Christopher Lean, Jonathan Livengood, Aidan Lyon, David Makinson, Daniel Malinksy, Hadrien Mamou, Stephen Mann, Alexandru Marcoci, Chris Marshall, Silvia Milano, James Nguyen, Deren Olgun, Philip Pettit, Miklós Rédei, Bryan Roberts, Joe Roussos, Thomas Rowe, Juha Saatsi, Alexander Sandgren, Jan Sprenger, Reuben Stern, Toby Solomon, Jeremy Strasser, Bastian Stuewer, Johanna Thoma, Aron Vallinder, Philippe van Basshuysen, Lachlan Walmsley, David Watson, Brad Weslake, Adam White, Timothy Williamson, James Willoughby, James Wills, Al Wilson, James Woodward, Nicolas Wüthrich, Shang Long Yeo, and Jiji Zhang.

I thank other friends in London for the kindness and camaraderie that they have shown to me over the years that I spent on this PhD, most notably Nik Simon, Raquel Plitt, Michael Naughton, David Watson, Eric Durell, Jonny Matthews, Rachel Muzyczka, Jamie Bender, Becky Waite, Dan Harper, Brita Cooper, Geoff Goodwin and Charlotte Bolland.

I thank my partner Julia Lefkowitz for all her love. Her support as I wrote this dissertation is just one of the uncountably many reasons that I am so grateful that she came into my life.

Above all, I thank my parents, Gita and Michael Kinney, and my brother Jonathan Kinney for their unconditional and constant love and encouragement throughout my life. For nearly thirty years, they have been my cheerleaders and my champions. I will never be able to put into words how much they mean to me.

This dissertation is dedicated to the memory of my grandmother, Phyllis Bellow.

Contents

List of Figures

List of Tables

0 Introduction

0.1 Granularity and Scientific Explanation

This dissertation is about the role of granularity in scientific explanation. When scientists explain some phenomenon, they can state their explanation at varying levels of granularity. Consider a prototypical scientific explanation: 'the patient developed lung cancer because they were a smoker'. One state of affairs—the patient being a smoker—explains another—the patient developing lung cancer. In the terminology of Hempel and Oppenheim (1948), the explanatory state of affairs (e.g. the patient being a smoker) is called the *explanans* and the state of affairs that is explained (e.g. the patient developing lung cancer) is called the *explanandum*. I will not use this terminology often in this dissertation, but occasionally it does allow for more straightforward expression of ideas.

On its face, it seems clear that any given explanandum can be explained using a more fine-grained or coarse-grained explanans. Instead of the explanation 'the patient developed lung cancer because they were a smoker', we can use the more fine-grained explanation 'the patient developed lung cancer because they were a Marlboro smoker', or the more coarse-grained explanation 'the patient developed lung cancer because they habitually inhaled carcinogens'. The first of these alternative explanations is more fine-grained because it contains more detail; it tells us not only that the patient smoked, but the brand of cigarette that they smoked. The second is more coarse-grained because it contains less detail; cigarette smoke is one of several carcinogens, and based on the coarse-grained explanation alone we do not know which of these carcinogens the patient habitually inhaled.

In light of these observations about the nature of explanation, it is clear that scientists face a problem. The same explanandum can be explained at multiple levels of granularity. Thus, we can ask whether there are principles for determining, for a given explanandum, what the optimal or correct level of granularity is for describing the explanans. Identifying and justifying such principles is the problem of granularity for scientific explanation, and the goal of this dissertation is to make progress on this problem.

In so doing, this dissertation contributes to an ongoing conversation in contemporary philosophy of science. I trace the origins of the current conversation to Hitchcock and Woodward (2003), who give an account of "explanatory depth", or the idea that some explanations of the same explanandum can be "deeper", or provide more understanding, depending on scientists' choice of explanans. Although Hitchcock and Woodward consider cases in which the difference in depth between various explanations is due largely to the different levels of granularity with which the explanans is described, Weslake (2010) argues that their theory of explanatory depth does not do enough to account for the fact that a coarse-grained explanation can sometimes be deeper than a more fine-grained one. Here, Weslake draws in part on the book-length treatment of explanatory depth given in Strevens (2008). Central to Weslake's critique of Hitchcock and Woodward is his claim that their account of explanatory depth is unable to recover the sense in which the "special sciences", i.e. those sciences other than fundamental physics, can provide better explanations of some events than fundamental physics. Similar worries motivate discussions of levels of granularity in scientific explanation by Potochnik (2010), Weatherson (2012), Franklin-Hall (2016), Clarke (2017), and List (forthcoming).

Discussions of the ideal level of granularity for a causal explanation are closely related to the debate surrounding *causal exclusion arguments* in philosophy of mind and metaphysics. Some proponents of causal exclusion arguments, such as Kim (1989, 2000) argue that causation can only occur at the level of fundamental physics. Therefore, causal explanations that cite a more coarse-grained explanans, such as the causal explanations offered by psychology, are elliptical and strictly false. Causal exclusion arguments are opposed by causal ecumenists such as Davidson (1970) and Jackson and Pettit (1988, 1990, 1990, 1992), who argue that genuine causal explanations of the same explanandum can be given at multiple levels of granularity. Most of the authors discussed in the previous paragraph are ecumenists who are concerned with finding an optimal level of granularity for causal explanation; they assume that causal explanation is possible when the explanans is described at different levels of granularity. By contrast, causal exclusionists assume that there is a *unique* level of granularity at which causal explanation is possible. The debate between exclusionists and ecumenists is part of the problem of granularity for scientific explanation, and will be discussed in detail in Chapter 1.

The remainder of this introduction proceeds as follows. First, I narrow the scope of my inquiry in this dissertation. I outline a graphical, interventionist theory of probabilistic causal explanation that I accept for the sake of my arguments; it is within the context of this theory that I aim to make progress on the problem of granularity. Second, I discuss the formal theory of random variables, and what it means, in a formal sense, for one random variable to be a coarsening or a refinement of another. This background will be important throughout the dissertation. Third, I

give an overview of the contents of each chapter. This overview shows how, over the course of this dissertation, I propose a decision-theoretic methodology for selecting an optimal level of granularity for describing the explanans in a given explanation. Fourth and finally, I offer a guide to the mathematical formalism and notation that I use throughout the dissertation.

0.2 Probabilistic Causal Explanation

The title of this dissertation correctly describes the problem of granularity as a problem for scientific explanation in general. However, the scope of my inquiry in this dissertation is in fact more limited. Specifically, I aim to make progress on the problem of granularity as it pertains to *probabilistic causal explanations* in the sciences. I make this decision for two reasons. First, many explanations in the sciences (arguably, the vast majority of them) are causal explanations. If we regard deterministic causal explanations as a special case of probabilistic explanations, then all causal explanations in the sciences fall under the category of probabilistic causal explanations. Second, graphical and interventionist theories of causation and causal explanation due to Pearl (2000), Spirtes et al. (2000), and Woodward (2003) provide a precise formal framework for describing causal explanations, both probabilistic and deterministic, in the sciences. This framework allows the problem of granularity to be precisely formulated.

The details of these accounts are described in Chapter 1. For the sake of this introduction, I need only discuss their most basic elements. A *Bayesian network* (or "Bayes net") consists of a causal graph and a probability distribution over the variables in that graph. A causal graph is a set of *random variables* and a set of ordered pairs of those variables (throughout this dissertation, 'variable(s)' will mean 'random variable(s)'). These ordered pairs are depicted graphically as arrows, or *directed edges* pointing from one variable to another. These directed edges represent a direct causal relationship between two variables, with the variable at the tail of the arrow representing the direct cause of the variable at the head of the arrow. Throughout this dissertation, I will assume for the sake of argument that any Bayes net under consideration accurately represents the causal structure of its target system in nature, unless I state otherwise. A more precise description of what it means for a Bayes net to accurately represent the causal structure of its target is provided in Chapter 1.

If a Bayesian network satisfies a formal axiom called the *Causal Markov Condition* (this too is discussed in more detail in Chapter 1), then we can calculate the probability distribution over the variables in a Bayes net, given an *intervention* on one

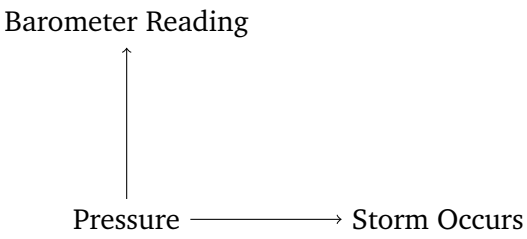

Fig. 0.1: Graphical Representation of a Storm System

or more variables in the graph. An intervention is understood as an exogenous change in the value of one or more variables in the graph, akin to an experimental intervention on a system in the practice of science. Thus, in a Bayes net that satisfies the Causal Markov Condition, one is able to determine which interventions change the probability distribution over other variables in the graph.

Following Woodward (2003), I hold that one variable is causally relevant to another just in case interventions changing the value of one variable lead to changes in the probability distribution over the value of another variable. This account of causal relevance is motivated by both the practice of experimental science and hypothetical reasoning. To illustrate how this works, consider a Bayesian network with three variables: one representing the readings of a barometer, one representing whether or not a storm occurs, and one representing the atmospheric pressure in a given region. The structure of this graph is shown in Figure 0.1. Pressure in the atmosphere, barometer readings, and the occurrence of storms will all be correlated. However, manually changing the reading on a barometer will not change the probability that a storm occurs. By contrast, if one could intervene on the atmosphere to change the pressure, then storms would become more likely. Thus, on an interventionist analysis, the barometer reading is not causally relevant to the occurrence of storms, whereas the pressure in the atmosphere is causally relevant to the occurrence of storms.

The interventionist theory of probabilistic causal relevance lays the groundwork for an interventionist theory of probabilistic causal explanation. Causal relevance is a relation that holds between variables; one variable or set of variables in a Bayes net can be causally relevant or not causally relevant to another variable or set of variables. By contrast, causal explanation is a relation that holds between the value taken by one variable and the value taken by another variable; one variable or set of variables takes a particular value or set of values because another variable or set of variables takes its value(s). Although my working theory of causal explanation is described in more detail in Chapter 1, the basic idea can be stated as follows. Let C and E variables in a Bayes net, let c_j be a value of C and let e_i be a value of E. The fact that $C = c_j$ causally explains the fact that $E = e_i$ if and only if

$p(e_i|do(c_j)) > p(e_i)$, where $do(c_j)$ represents the fact that the variable C is set to the value c_j via an intervention. For example, in the case described above, interventions setting the pressure to lower values will increase the probability of a storm occurring, where an increase is defined in relation to the prior probability of a storm occurring. Thus, the atmospheric pressure being at some low level can causally explain why a storm occurs. By contrast, there is no intervention on the value of the barometer reading that increases the probability that a storm occurs, and so no barometer reading can causally explain why a storm occurs.

In summary, my approach to probabilistic causal explanation in this dissertation can be described as follows. Bayesian networks are used to represent the causal structure of systems in nature. These Bayes nets can be used to calculate the probability distribution over variables in the graph, given other interventions on other variables in the graph. These interventional probability distributions allow us to determine which variables in the Bayes net are causally relevant to other variables in the Bayes net. The interventional probability distributions also allow us to determine, for any variable E that takes the value e_i, which values taken by other variables explain why it is the case that $E = e_i$.

0.3 Random Variables and Granularity

In the previous sections, I make reference to random variables and the probability distributions defined over them. In this section, I describe these concepts in a precise way, and show how this formal description can be used to give a clear statement of the problem of granularity. Let us begin with the measure-theoretic definition of a random variable X. Let $(R_X, \mathcal{R}_X)$ be a *measurable space*, where R_X is a set, and $\mathcal{R}_X$ is an algebra on R_X. This means that $\mathcal{R}_X$ is a collection of subsets of R_X that includes R_X and is closed under complement, union, and intersection. An $(R_X, \mathcal{R}_X)$-valued random variable X is a function $\Omega \to R_X$. The domain Ω is called the *outcome set*, and will be taken here to represent a set of possible worlds. Throughout this dissertation, I will assume that all random variables are defined over the same outcome set (in Chapter 5, I will defend this assumption when discussing alternative approaches to coarsening). Following convention, I refer to the elements of the set R_X as the *values* of X, and R_X as the *value set* of X. Since X is a function, each element of Ω is mapped to only one element of R_X, although multiple elements of Ω can be mapped to the same element of R_X. R_X can be interpreted as a set of mutually exclusive possible ways that a given world might be. Note that although X is a function on the set of possible worlds, when X takes some value x_i, we usually write this as $X = x_i$, rather than $X(\omega) = x_i$, where ω is some possible world. This is because the possible world that is mapped to x_i is usually clear from the context

in which the variable is used (i.e. it is the actual world or some clearly specified counterfactual possible world).

To illustrate, suppose that I toss a coin, and that a random variable $X : \Omega \rightarrow R_X$ represents the outcome of the coin toss. The value set R_X is defined as the set $\{x_1, x_2\}$, where x_1 represents the coin landing heads, and x_2 represents the coin landing tails. We assume that in all worlds in Ω, the coin is tossed and lands either heads or tails. Thus, the random variable X groups a set of possible worlds into those where the coin lands heads and those where the coin lands tails.

A probability space is an ordered triple $\langle \Omega, \mathcal{A}_\Omega, p(\cdot) \rangle$, where the set Ω is again the outcome set, the set $\mathcal{A}_\Omega$ is an algebra on Ω, and $p(\cdot)$ is a *probability distribution* over $\mathcal{A}_\Omega$. This means that $p(\cdot)$ is a function $\mathcal{A}_\Omega \rightarrow [0, 1]$ that satisfies Kolmogorov's (1933) three probability axioms:

1. For all $D \in \mathcal{A}_\Omega$, $p(D) \geq 0$.

2. $p(\Omega) = 1$.

3. For all pairs $(D \in \mathcal{A}_\Omega, F \in \mathcal{A}_\Omega)$ where D and F are disjoint, $p(\{D, F\}) = p(D) + p(F)$.

These axioms just say that probabilities are always non-negative, that the probability of the union of all possibilities is 1, and that the probability of the union of any two disjoint sets of sets of possible worlds is equal to the sum of the probabilities of each set in the union.

A random variable X is measurable with respect to a probability space $\langle \Omega, \mathcal{A}_\Omega, p(\cdot) \rangle$ if and only if, for all $Z \in \mathcal{R}_X$, $\{\omega : X(\omega) \in Z\} \in \mathcal{A}_\Omega$. If X is measurable with respect to a probability space $\langle \Omega, \mathcal{A}_\Omega, p(\cdot) \rangle$ and for every subset $S \subseteq R_X$, $X^{-1}(S) \in \mathcal{A}_\Omega$, then we can assign a probability to each value of X, as well as each negation or disjunction of value(s) of X. To see why, let S be a (possibly singleton) set containing some value(s) of X. If S is a singleton, then S contains the value taken by X. If S is not a singleton, then S contains the values of X consistent with a negation or disjunction of some value(s) of X. Since $X^{-1}(S) \in \mathcal{A}_\Omega$, we can define a pre-image $X^{-1}(S) = \{\omega : X(\omega) = S\} \in \mathcal{A}_\Omega$, which means that $X^{-1}(S)$ is assigned a probability by the function $p(\cdot)$. The probability $p(X^{-1}(S))$ is the probability assigned to the value of X, or to the negation or disjunction of value(s) of X, that was identified with $S \subseteq R_X$. Throughout the dissertation, I assume that all random variables are measurable with respect to a common probability space $\langle \Omega, \mathcal{A}_\Omega, p(\cdot) \rangle$, and that for any random variable X used herein, $S \subseteq R_X$, $X^{-1}(S) \in \mathcal{A}_\Omega$ (again, I

will defend this assumption in Chapter 5 when discussing alternative approaches to coarsening).

As a brief aside, I note here that throughout the dissertation I assume that all random variables have countably many elements in their value space. This means that, for any two variables X and Y with values x_i and y_j, the conditional probability $p(x_i|y_j)$ can be defined using the ratio formula:

$$p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)} \text{ where } p(y_j) \neq 0. \tag{0.1}$$

The ratio formula can also be written as Bayes' theorem:

$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{p(y_j)} \text{ where } p(y_j) \neq 0. \tag{0.2}$$

Where the conditioning event of a conditional probability has a prior probability equal to zero, that conditional probability is undefined.

We are now in a position to clarify how random variables can be used to state the problem of granularity in a more precise way. Let X' be a random variable. X is a *coarsening* of X' if and only if the value set R_X is a *quotient set* of the value set $R_{X'}$, and R_X has less elements than $R_{X'}$. A quotient set is defined as follows. Let $\sim$ be an equivalence relation (i.e. a symmetric, reflexive, and transitive binary relation) defined over the elements of $R_{X'}$. An equivalence class in $R_{X'}$ according to $\sim$ is a set of elements of $R_{X'}$ that all stand in the relation $\sim$ to one another. A quotient set on $R_{X'}$ according to the relation $\sim$ is the set of all equivalence classes of $R_{X'}$. Thus, X is a coarsening of X' if and only if the value set R_X is the set of all equivalence classes, according to some equivalence relation, in the value set $R_{X'}$. If X is a coarsening of X', then X' is a *refinement* of X.

Note that when discussing the values of a coarsening X of a more fine-grained variable X', I will denote these coarse-grained values using single letters, rather than using sets of values of the fine-grained variable. For instance, suppose that there is an equivalence class $\{x'_i, x'_j\}$ in the value set of X', according to an equivalence relation $\sim$. If X is a coarsening of X' according to $\sim$, then $\{x'_i, x'_j\}$ is an element of the value set of X. However, if X takes this value, then the fact that X takes this value may be denoted by writing $X = x_k$, where it is implicit that $x_k = \{x'_i, x'_j\}$.

To illustrate, let X' be a random variable with the value set $\{2, 3, 4, 5\}$. Let $\sim$ be an equivalence relation on the value set of X' such that two elements are equivalent if and only if they are both even or both odd. Coarsening according to this equivalence relation yields the random variable X, with the value set $\{\{2, 4\}, \{3, 5\}\}$. This can then be re-written as the value set $\{\text{even}, \text{odd}\}$.

The problem of granularity can now be stated as follows. According to a Bayesian network that adequately represents a given system, the fact that a variable E takes some value is causally explained by the fact that another variable C takes some value. However, in many cases it appears that we can replace C with a refinement or a coarsening of C in a way that leads to a more fine-grained or coarse-grained explanans, while preserving the explanatory relationship between cause and effect. This gives rise to several questions:

- For any given explanandum, is there a correct level of granularity for the value set of the explanans variable?

- What is the optimal level of granularity with which we should define an explanans variable in a Bayesian network?

- If an optimal level of granularity for an explanans variable exists, which equivalence class can we define on more fine-grained variables so as to produce a coarsening at the optimal level of granularity?

These questions, and others like them, together comprise the problem of granularity for scientific explanation when the domain of scientific explanations is limited to those explanations that follow the graphical and interventionist model of probabilistic causal explanation.

Stated in this way, the problem of granularity is an important subset of a larger problem within graphical and interventionist causal explanation, namely the so-called *problem of variable choice*. In a paper which does not claim to solve the problem of variable choice, but does give an overview of its scope and difficulty, as well as some potential avenues for solutions, Woodward (2016) describes this problem as follows:

> Suppose we are in a situation in which we can construct or define new previously unconsidered variables either de novo or by transforming or combining or aggregating old variables, and where our goal is to find variables that are best or most perspicuous from the point of view of causal analysis/explanation. Is there anything useful to be said about the considerations that ought to guide such choices? (p. 1048).

In the context of Woodward's quote, coarsening or refining variables in a causal model amounts to "transforming" existing variables to define new variables. As such, the problem under consideration in this dissertation is a subset of the wider problem that he identifies. However, while Woodward's paper on the wider problem of variable choice mostly gives the lay of the land, without endorsing a particular

positive proposal, my goal in this dissertation is to make progress on the narrower problem of granularity.

0.4 A Preview of the Chapters

As mentioned earlier in this introduction, Chapter 1 discusses the debate between proponents of various causal exclusion arguments (i.e. those who claim that for any given event, there is a unique level of granularity at which we can give a causal explanation of that event) and causal ecumenists (i.e. those who claim that the same event or phenomenon can often be causally explained at multiple levels of granularity). I argue that the Bayesian network approach to representing the causal structure of systems in nature is consistent with causal ecumenism, considering and rebutting specific exclusion arguments due to Gebharter (2017) and List and Menzies (2009). Thus, I provide a definitive answer to the first question that I identify above as part of the problem of granularity for scientific explanation. That is, I show that within the context of the graphical and interventionist approach to causal explanation, a given explanandum can be explained by an explanans variable defined at multiple levels of granularity.

Having defended an ecumenist answer to the first question posed above, I turn to the task of identifying the optimal level of granularity for an explanans variable, given some explanandum. In Chapter 2, I explore whether various Bayesian measures of explanatory power proposed by Schupbach and Sprenger (2011), Crupi and Tentori (2012), Ströing (2018), and Eva and Stern (forthcoming) can help answer this question. My finding is largely negative. The measures proposed by these authors necessarily track the degree to which the probabilistic relationship between an explanatory hypothesis and the evidence that the hypothesis explains allows an agent to accurately predict whether some evidence is observed, conditional on the hypothesis being true. However, scientific judgments about whether some hypothesis is the best or deepest explanation of some evidence do not always track predictive accuracy in the same way. This fact about scientific practice becomes apparent when we assess the comparative goodness of explanations that differ only with respect to the granularity with which the hypothesis is stated, i.e. the class of cases that are relevant to the problem of granularity. This conclusion stands in contrast with some claims by Schupbach and Sprenger, who hold that such measures do allow us to assess the overall goodness of an explanation.

Chapter 3 lays out the central positive proposal of this dissertation. I begin by considering the thesis that a more proportional relationship between a cause and its effect yields a more abstract causal explanation of that effect, thereby producing

a deeper explanation. This thesis, defended by Weslake (2010) and formalized by Pocheville et al. (2017), has important implications for how we choose the optimal granularity of explanation for a given explanandum. In this chapter, I argue that this thesis is not generally true of probabilistic causal relationships. In light of this finding, I propose a pragmatic measure of explanatory depth. This measure uses a decision-theoretic model of information pricing to determine the optimal granularity of explanation for a given explanandum, agent, and decision problem. In a nutshell, my thesis is that we can find the optimal level of granularity for a given causal variable by finding the coarsest possible causal variable such that an agent with a utility function defined over the effect variable would pay just as much to learn the value of the coarse-grained variable as they would pay to learn the value of the refinements of that variable.

Thus, I define the optimal level of granularity for a causal variable in a way that is fundamentally relative to the interests of a particular agent. This explicitly pragmatic approach is a departure from previous approaches to the question of explanatory depth or goodness as it pertains to granularity; most authors in these debates insist on an interest-neutral conception of what counts as the optimal level of granularity for a causal variable. In Chapter 3, I defend my decision to break with this tradition, and argue that embracing the idea that explanatory goodness has an ineliminable pragmatic component allows us to make progress on the problem of granularity for scientific explanation.

Chapter 4 applies the pragmatic approach to coarsening put forward in Chapter 3 to contemporary research in machine learning. Specifically, it both builds on and critiques Chalupka et al.'s (2015, 2016, 2016, 2017) framework for learning coarse-grained variables from more fine-grained data; to my knowledge, this is the main contribution in the machine learning literature to coarse-graining in probabilistic causal models. I outline Chalupka et al.'s proposal and highlight its positive features, making it clear that it would be a bad result for my pragmatic approach to coarsening if it could not recover these same features. However, I argue that my approach can recover these features, such that Chalupka et al.'s framework is not a threat to the viability of my account. In making this argument, I extend my formal model for finding the optimal coarsening of variables in a causal model to effect variables; up to this point in the dissertation, I had only considered coarsenings of the causal variable.

The discussion of coarsening and refining in the context of probabilistic models invites discussion of another ongoing debate in recent philosophy of science, philosophy of probability, and metaphysics. This is the debate over whether probabilistic relationships at higher levels of description can be *objective chances*, even if the most accurate model of nature at a finer level of description turns out to be entirely

deterministic. I call the thesis that there are higher-level chances despite lower-level determinism the *emergent chance thesis*. In Chapter 5, I turn my attention to this debate, focusing specifically on an argument in favor of the emergent chance thesis put forward by List and Pivato (2015b). While I do not come to an all-things-considered conclusion on the truth of the emergent chance thesis, I do argue that List and Pivato's argument fails to establish its truth. This clarifies that I do not necessarily take my arguments in previous chapters for the occasional optimality of coarse-grained, probabilistic, causal explanations to be arguments in favor of the emergent chance thesis.

Finally, Chapter 6 is a slight departure from the main theme of the dissertation. Instead of considering coarsenings of variables in a causal model, this chapter considers coarsenings of the probability distribution over the variables in a graphical causal model. Specifically, instead of a single probability distribution over the variables in a causal model, a set of probability distributions is used to model the relationships between the same variables. These imprecise probabilities have been a topic of extensive study by formal epistemologists—see Bradley (2016) for a comprehensive review—but these discussions have not touched on probabilistic causal models. By contrast, causal models that use imprecise probabilities have been studied by computer scientists—see Corani et al. (2012)—but these authors do not pay much philosophical attention to the question of how imprecise probabilities can be used to represent the causal structure of systems in nature. In this chapter, I conclude that there are serious limitations to the representative power of causal models that use a set of probabilities rather than a single probability distribution.

Although these chapters were originally written as stand-alone papers, I have edited them so that they make reference to each other where necessary, and so that they read as a coherent whole. At the same time, some technical material and especially important concepts are repeated, so that the reader generally does not have to remember content from previous chapters in order to understand each chapter. The exception to this rule is Chapter 4, which builds directly on Chapter 3 and therefore refers to it frequently. Each chapter concludes with an appendix in which mathematical propositions are proved; there are no mathematical proofs in the main text of any of the chapters.

0.5 Guide to Formalism

I conclude this introduction with a guide to some of the mathematical formalism used throughout this dissertation. While this guide does not cover every single

formal symbol used in the dissertation, it will hopefully be helpful to the reader over the course of the dissertation.

Random variables are represented using capital letters, e.g. the random variable X. Where multiple random variables must be listed, superscript numbers are used, e.g. the random variables $V^1, V^2, ..., V^n$. Superscripts are also used in several chapters to indicate refinements and coarsenings of a given variable. Values of random variables are represented using lower-case letters with a subscript letter, i.e. x_i. The letter used for variable values should always match the letter used to denote the random variable, e.g. the random variable X has the value x_i. The subscript letter is replaced with a subscript number in cases where values of a random variable need to be listed, e.g. the values x_1, x_2,..., x_n. Occasionally, more evocative values for random variables, such as words or numbers, are used. These cases should be readily comprehensible to the reader in light of the context in which they appear.

More complex groupings of random variables, such as sets of random variables, relations defined over random variables, graphs and Bayesian networks are represented using calligraphic script, e.g. the variable set $\mathcal{V}$. When representing the n-tuple of values taken by a group of variables, bold lower-case letters with subscripts are used, e.g. the variables in $\mathcal{V}$ take the values in the n-tuple $\mathbf{v_i}$.

Functions other than random variables are denoted by lower case letters, followed by the symbol $(\cdot)$, e.g. the probability function $p(\cdot)$. When denoting the probability that a random variable takes a particular value, the random variable itself is usually omitted, e.g. $p(x_i)$ is written instead of $p(X = x_i)$. Where the random variable is not omitted, it is included for the sake of comprehension.

In Chapters 2 and 5, I sometimes discuss probabilities that are assigned to elements of an algebra of possible worlds, rather than the values of random variables. In these contexts, I use capital letters to represent these elements of an algebra, and calligraphic capital letters to represent the algebra, with a subscript to indicate the set of possible worlds over which the algebra is defined, e.g. the element Z of the algebra $\mathcal{A}_\Omega$.

1 Bayesian Networks and Causal Ecumenism

1.1 Introduction

Suppose that I form the desire to move my arm, and subsequently move my arm. There are at least two ways of describing this process. On the one hand, a mental event (I form the desire) causes an action (I move my arm). On the other hand, a neural event (a physical process occurs in my brain that accounts for my forming the desire) causes me to move my arm. This multi-level picture is not unique to the relationship between psychology and neurobiology. One could tell a similar story with respect to the relationship between economics and psychology, biology and chemistry, chemistry and physics, and so on. Putative multi-level causal relations can cross disciplines: one could tell a coarse-grained story about El Niño weather patterns having a particular effect on the GDP of Australia, or one could tell a fine-grained story about certain combinations of sea surface temperatures and zonal winds having the same impact. In each instance, the level of causal description that we give for an event corresponds to the level at which we seek to explain that same event. When we ask why I moved my arm, or why the growth rate of Australian GDP slowed, it seems that we can give causal explanations at different levels of granularity. As discussed in the introduction to this dissertation, the thesis that one can often give an adequate causal explanation of the same event at varying levels of granularity is defended by Jackson and Pettit (1988, 1990, 1990, 1992, see also Pettit 2017). This thesis is also implicit in Davidson's (1970) philosophy of mind. Following Jackson and Pettit, I label this view *causal ecumenism*.

I am sympathetic to Hitchcock's (2012) argument that since the soundness of any causal exclusion argument is sensitive to one's favored of theory of causation, we should be hesitant to take any argument for or against causal exclusion to be generally decisive. Rather, it seems that the best that we can do is determine whether or not a causal exclusion argument is sound within the context of a particular theory of causation. It is this more modest task that I take on in this chapter. In both philosophy of science and the sciences themselves, *Bayesian Networks* (or "Bayes nets")—as developed by Pearl (2000) and Spirtes et al. (2000)—are a powerful formalism for representing causal structure. Bayes nets also entail a semantics

for determining when one event causally explains another. This semantics makes essential use of counterfactual conditionals about hypothetical interventions on the causal structure of the system described by a given Bayes nets. These interventional conditionals are at the core of Woodward's (2003) highly influential account of causal explanation. In this chapter, I argue that the Bayes nets formalism, and the interventionist semantics for causal explanation that it entails, is in keeping with Jackon-and-Pettit-style causal ecumenism. My conclusions are similar to those put forward in Eronen (2012), Woodward (2015a), and Polger et al. (2018). However, these authors only consider the causal exclusion problem insofar as it pertains to Woodward's interventionist semantics in deterministic cases; they do not frame their arguments in terms of the Bayes nets formalism, nor do they discuss probabilistic cases. Thus, my conclusions here have a wider scope than these earlier arguments.

There are those that disagree with this ecumenical approach. Broadly speaking, there are two species of *causal exclusion* arguments against the view described above. The first, which is most strongly associated with Kim (1989, 2000), holds that all genuine causation occurs at the most ontologically fundamental level of description. On this view, all higher-level properties cited in causal explanations must be epiphenomenal, since denying this epiphenomenalism would lead to the putatively false result that all physical explananda are causally over-determined. Thus, the existence of fine-grained causal explanations results in the *upward exclusion* of more coarse-grained explanations. I will not address the metaphysical aspects of this argument here, but focus instead on a recent revival of Kim's approach by Gebharter (2017). Gebharter argues that the Bayes nets approach is consistent with upward exclusion in all cases. His argument rests on the result that models that represent both higher- and lower-level properties of the same event fail to satisfy axioms of the Bayes nets formalism.

The second version of causal exclusion, which is advocated by List and Menzies (2009), holds that there may be an optimal granularity of description for any given phenomenon, but that this optimal level need not be maximally fine-grained or ontologically fundamental. Whether there is an optimal level of causal explanation depends on contingent features of a causal system. Where such a level exists, it is at this optimal level, and only at this level, that causal explanation can occur. Where a coarser-grained level of description explains some phenomenon, List and Menzies hold that finer-grained descriptions fail to be explanatory. To illustrate, they argue that if the explanation 'the ball hit the window with sufficient velocity, causing it to break', explains why some window broke, then a molecular description of the ball hitting the window would fail to explain why the window broke. They call this type of exclusion *downward exclusion*. On their view, whether or not there is downward exclusion depends on the relationship between the explanatory proposition or event and the proposition or event to be explained. Namely, the two

relata of the explanation must be *proportional* to one another, in a sense that will be made precise in what follows. Importantly, List and Menzies' argument for causal exclusion hinges on contingent and *a posteriori* features of the relation between cause and effect. This stands in contrast to Kim-style causal exclusion arguments, which begin with premises about the *a priori* nature of causation in general.

Here is the plan for this chapter. In Section 1.2, I introduce the Bayes nets approach to representing the causal structure of systems, and detail how the Bayes nets axioms imply an interventionist semantics for causal explanation. In Section 1.3, I argue that in the Bayes nets context, the causal exclusion problem is a problem of variable choice, and that the Bayes nets formalism supports a generally ecumenical response to causal exclusion arguments. In Section 1.4, I respond to Gebharter's argument that the Bayes nets formalism entails upward causal exclusion. In Section 1.5, I respond to List and Menzies' argument that a difference-making approach to causal semantics entails downward exclusion in many cases. In Section 1.6, I offer concluding remarks.

1.2 Background

1.2.1 Bayesian Networks

A Bayes net is a triple $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$. $\mathcal{V}$ is a set of random variables whose values denote different possible states of the system being represented. $\mathcal{E}$ is an acyclic set of ordered pairs of the variables in $\mathcal{V}$. These ordered pairs, called *edges*, are usually represented visually as arrows pointing from one variable to another. The pair $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is called a *graph*. If a graph contains an edge $V^1 \rightarrow V^2$, then V^1 is a *parent* of V^2, and V^2 is a *child* of V^1. A chain of parent-child relationships is called a *directed path*. If there is a directed path from a variable V^3 to a variable V^4, then V^4 is a *descendant* of V^3 and V^3 is an *ancestor* of V^4. Finally, $p(\cdot)$ is a probability distribution defined over the Cartesian product of the value sets of each variable in $\mathcal{V}$. In any Bayes net, this probability distribution satisfies the following axiom:

> **Markov Condition:** Conditional on its parents, any variable in a Bayes net is probabilistically independent of its non-descendants.

This property of the probability distribution in a Bayes net plays a crucial role in justifying the use of Bayes nets to represent the causal structure of systems. Following Spirtes et al. (2000, p. 475), let us stipulate a variable set $\mathcal{V}$ is *causally sufficient* if and only if changes in the world that are not represented in the model and are common causes of two or more variables in the model do not make a difference to

the probability distribution over $\mathcal{V}$. Spirtes et al. (2000, p. 53) then introduce the *causal Markov condition* for a Bayes net, which is defined as follows:

> **Causal Markov Condition (CMC):** If $\mathcal{V}$ is causally sufficient and $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ represents the true causal structure of its target system, then the observed probability distribution over $\mathcal{V}$ satisfies the Markov Condition.

Throughout this dissertation, I will often say that a Bayes net "satisfies CMC". I intend this to mean that a Bayes net satisfies the Markov Condition, under the supposition that the Bayes net's variable set is causally sufficient and its graph represents the true causal structure of the system.

In what follows, I will stipulate that certain Bayes nets are *adequate representations* of target systems. This notion of adequate representation can be made precise as follows. Let $\mathcal{V}_\Phi$ be a variable set such that the values of each variable $X \in \mathcal{V}_\Phi$ represent a different state or set of states of some system Φ. Let $p_\Phi(\cdot)$ be the probability distribution over the values of those variables that has maximum likelihood, given the observed data. Let us suppose further that the variable set $\mathcal{V}_\Phi$ is sufficient to represent the causal structure of the system in question. A Bayes net $\mathcal{N}_\Phi = \langle \mathcal{V}_\Phi, \mathcal{E}_\Phi, p_\Phi(\cdot) \rangle$ is an adequate representation of Φ just in case it satisfies CMC and Minimality and its graphical structure matches the true causal structure of Φ. Determining the true causal structure of some system is difficult. Given a variable set and a probability distribution over the values of those variables, standard causal modeling algorithms, such as those developed by Spirtes et al. (2000), often return more than one Bayes net that satisfies CMC and Minimality, with edges potentially pointing in different directions. Experiments can be helpful in ruling out certain graphical structures, so as to determine a single Bayes net that is an adequate representation of a target system. Additionally, sophisticated causal modeling techniques developed by Shimizu et al. (2006) and Janzing and Scholkopf (2010) may, in some cases, allow for the inference of a single representationally adequate causal graph from purely observational data. Going forward, I will make the idealizing assumption that for a given target system Φ, scientists are capable of determining, for a given variable set $\mathcal{V}_\Phi$ and joint probability distribution $p_\Phi(\cdot)$, a single causal graph that is an adequate representation of Φ.

Throughout this dissertation, I take the edges of a Bayes net to represent a relation of direct causal relevance between the variables that are related by an edge. To illustrate, suppose that a Bayes net contains an edge $X \rightarrow Y$. This is meant to represent the fact that if $X = x_i$ and $Y = y_j$, then the event represented by X taking the value x_i is directly causally relevant to the event represented by Y taking the value y_j. This discussion raises a seemingly pertinent question: what exactly is meant by a relation of direct causal relevance? The theory of Bayes nets does not

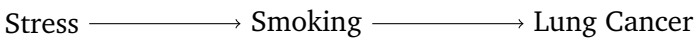

Fig. 1.1: Simple Causal Graph

answer this question by providing a set of necessary and sufficient conditions for one event to be a direct cause of another. Instead, the use of Bayes nets to represent causal structure depends on a "non-reductive" or "Euclidean" approach to clarifying the nature of causation (see Hausman and Woodward (1999), Spirtes et al. (2000, p. 3)). Rather than providing necessary and sufficient conditions for X to be directly causally relevant to Y, it is argued that CMC, along with a Minimality condition to be introduced below, supports a causal interpretation of the Bayes net. In other words, it is argued that if a graph $\mathcal{G}$ adequately represents the causal structure of its target system, then we should expect the observed probability distribution over $\mathcal{V}$ to satisfy CMC and Minimality with respect to to $\mathcal{G}$.

Why should we expect the observed probability distribution over a representationally adequate causal graph to satisfy CMC and Minimality with respect to that graph? I begin my answer to this question by noting that CMC has two crucial implications that support the causal interpretation of a Bayes net. First, any probability distribution $p(\cdot)$ that satisfies CMC with respect to some graph $\mathcal{G}$ will also satisfy the *Screening-Off Condition*, defined as follows.

> **Screening Off:** For any variable X, if $\mathcal{PA}_X$ is the set containing all the parents of X and W is an ancestor of some variable in $\mathcal{PA}_X$, then X and W are independent, conditional on $\mathcal{PA}_X$.

To illustrate, consider the simple causal graph in Figure 1.1, in which the variable 'Stress' represents a person's level of stress, 'Smoking' represents the number of cigarettes that a person smokes, and 'Lung Cancer' is a binary variable that represents whether or not a person develops lung cancer. Screening Off says that once we know the value of the variable Smoking, the probability that the person develops lung cancer is independent of the value of Stress. More generally, CMC ensures that once we have full information about all the direct causes of an event, no additional information about the causes of those direct causes should change the probability that we assign to an event.

The second important corollary of CMC is closely related to Reichenbach's (1956) *Common Cause Condition*, and is defined as follows.

> **Common Cause:** For any disjoint sets of variables $\mathcal{X}$ and $\mathcal{Y}$, if $\mathcal{X}$ and $\mathcal{Y}$ are not unconditionally independent, then every pair of subsets $\mathcal{X}^{\dagger} \subseteq \mathcal{X}$

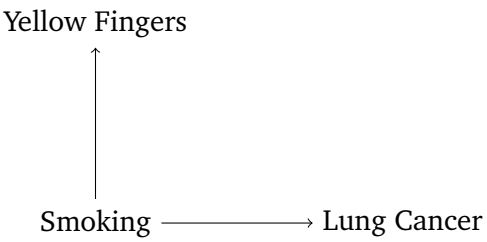

Fig. 1.2: Correlated Variables with Common Cause

> and $\mathcal{Y}^\dagger \subseteq \mathcal{Y}$ are such that that either at least one variable in $\mathcal{X}^\dagger$ is an ancestor or descendant of at least one variable in $\mathcal{Y}^\dagger$, or $\mathcal{X}^\dagger$ and $\mathcal{Y}^\dagger$ are independent conditional on a (possibly empty) set $\mathcal{A}_{\mathcal{X}^\dagger,\mathcal{Y}^\dagger}$ of variables that are a common ancestor of at least one variable in $\mathcal{X}^\dagger$ and one variable in $\mathcal{Y}^\dagger$.

To illustrate, consider the graph in Figure 1.2, in which the variables 'Smoking' and 'Lung Cancer' have the same meaning as they do in Figure 1.1, while the variable 'Yellow Fingers' represents, on some scale, how yellow a person's fingers are. Suppose that there is an observed correlation between Yellow Fingers and Lung Cancer; unconditionally, the two variables are not independent. If, as is likely the case, there is no direct causal link in either direction between Yellow Fingers and Lung Cancer, Common Cause tells us that the two variables must share a common ancestor, namely Smoking, on which they are conditionally independent. This shows that since a Bayes net satisfies CMC, then any correlations between variables in that Bayes net are accounted for, either via a directed path from one variable to the other, or via a common ancestor of the correlated variables.

Crucially, we can show that the following proposition is true:

Proposition 1. *A probability distribution $p(\cdot)$ satisfies CMC with respect to $\mathcal{G}$ if and only if it satisfies Screening Off and Common Cause with respect to $\mathcal{G}$.*

Thus, Screening-Off and Common Cause are both necessary conditions for the causal interpretation of a Bayes net if and only if CMC is also a necessary condition for the causal interpretation of a Bayes net. The relationship between CMC, Screening Off and Common Cause has been discussed throughout the causal modeling literature (e.g. Williamson 2005), so that the proof of Proposition 1 is not especially novel. However, to my knowledge the biconditional statement of the relationship between the three conditions in Proposition 1 is not explicitly stated elsewhere. In what follows, I will assume that CMC is indeed a necessary condition for the causal interpretation of a Bayes net.

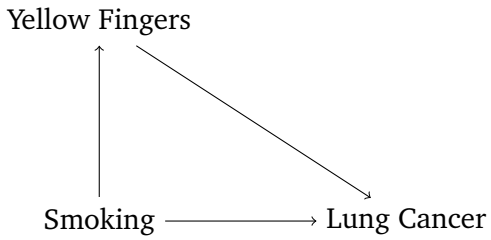

Fig. 1.3: Graph with Excess Edge

However, CMC cannot be the sole axiom justifying a causal interpretation of a Bayes net. This is because CMC does nothing to restrict the number of edges that can be included in a graph. To illustrate, consider Figure 1.3. Here, as in Figure 1.2, let us assume that Yellow Fingers and Lung Cancer are independent conditional on Smoking. The inclusion of an edge between Yellow Fingers and Lung Cancer, in spite of the fact that they are independent conditional on Smoking, does not induce a violation of CMC. So CMC alone does not rule out a Bayes net that represents a scenario in which having yellow fingers causes lung cancer, even when the correlation between these two variables is entirely explained by the increased presence of smoking in those with both yellow fingers and lung cancer. Thus, we need a further constraint on Bayes nets to ensure that they do not contain excess edges, i.e. an Ockham's razor condition to ensure that Bayes nets have the simplest structure needed to account for the observed data.

One axiom that is designed to ensure that a Bayes net does not contain excess edges is the *Minimality Condition*, which is defined as follows.

> **Minimality:** A Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ satisfies CMC, and any Bayes net $\mathcal{N}^* = \langle \mathcal{V}, \mathcal{E}^*, p(\cdot) \rangle$ such that $\mathcal{E}^* \subset \mathcal{E}$ does not satisfy CMC.

The graph in Figure 1.3 violates Minimality, since removing the arrow between Yellow Fingers and Lung Cancer does not result in a violation of CMC. In general, Minimality ensures that the edges in a Bayes net are only doing the work required in order to satisfy CMC. This axiom will be crucial to my response to Gebharter's exclusion argument in Section 1.4.[1]

[1]Many readers will be familiar with the Faithfulness condition for Bayes nets, which is strictly stronger than Minimality. I choose Minimality over Faithfulness as an adequacy condition for the causal interpretation of Bayes nets, since there is a case to be made that Bayes nets that satisfy Minimality but not Faithfulness are accurate representations of some causal systems. For a perspicuous comparison of the two conditions, see Zhang (2012).

1.2.2 Intervention and Causal Explanation

The Bayes nets approach to causal modeling entails a semantics of causal explanation in terms of interventional counterfactual conditionals. An *intervention* on a Bayes net is an exogenous change in the value(s) of a variable or group of variables in that Bayes net. By an 'exogenous change', I mean a change that occurs independently of the values taken by any other variables in the Bayes net, as though the change were achieved by a manipulation of the target system by factors external to the system.

Woodward (2003, p. 203) formalizes the connection between intervention and causal explanation as follows. Let $Y = y_j$ represent the event to be explained. Let the explanatory propositions include the event represented by $X = x_i$, as well as some theoretical generalizations T describing the relationship between the variables X and Y. He claims that the following three conditions are individually necessary and jointly sufficient for an explanation of why $Y = y_j$:

(i) Propositions expressing that $X = x_i$ and that $Y = y_j$ are true or approximately true, as are the generalizations in T.

(ii) According to T, if X is set to x_i via an intervention, then $Y = y_j$.

(iii) There is some intervention setting X to x_k, where $x_i \neq x_k$, such that, according to T, $Y = y_l$, where $y_j \neq y_l$.

In a Bayes nets context, the theoretical generalizations T describing the relationship between X and Y are propositions describing the structure of the graph and the joint probability distribution over the graph. Thus, (i) can be straightforwardly adapted into a Bayes nets context.

However, (ii) and (iii) need to be translated into probabilistic language. I propose that (ii) and (iii) can be combined into a single condition which, together with a straightforward translation of (i) into the Bayes nets context, are individually necessary and jointly sufficient conditions for probabilistic causal explanation. Thus, Woodward's conditions for causal explanation can be translated into probabilistic language as follows:

(i*) Propositions expressing that $X = x_i$ and that $Y = y_j$ are true or approximately true, and there is a Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ that is an adequate representation of a given system, and is such that $X \in \mathcal{V}$ and $Y \in \mathcal{V}$.[2]

[2] I include the qualification "approximately true" here to follow Woodward; I do not wish to commit myself to any particular view of approximate truth.

(ii*) According to the joint distribution $p(\cdot)$, $p(y_j|do(x_i)) > p(y_j)$.

The $do(\cdot)$ operator denotes that the variable X is set to the value x_i via an intervention. The condition (ii*) replaces the implication relation in (ii) with a probability-raising relation. Performing a similar translation on condition (iii) would yield the condition that there is some pair of values (x_k, y_l) such that $p(y_l|do(x_k)) > p(y_l)$. The truth of (ii*) implies that such a pair exists, due to the following proposition:

Proposition 2. *If there is a Bayes net $\mathcal{N}$ with variables X and Y, with values x_i and y_j such that $p(y_j|do(x_i)) > p(y_j)$, then there is some pair of values (x_k, y_l) such that $x_i \neq x_k$, $y_j \neq y_l$, and $p(y_l|do(x_k)) > p(y_l)$.*

Thus, it is possible to translate Woodward's deterministic account of causal explanation, defined in terms of interventional counterfactual conditionals, into a probabilistic, interventionist account of causal explanation.[3]

The approach to causal explanation presented here is contrastive. The relation of probability raising under intervention between $X = x_i$ and $Y = y_j$ not only implies that the occurrence of the event represented by $X = x_i$ explains why the event represented by $Y = y_j$ occurs. It also implies that the occurrence of the event represented by $X = x_i$ explains why the event represented by $Y = y_j$ occurs *rather than* the set of events represented by $Y \neq y_j$. This is a special case of contrastive explanation, one in which the "fact" to be explained, i.e. the event represented by $Y = y_j$, is contrasted with a "foil" that is its logical negation.[4] This positive instance of contrastive explanation is underwritten by the fact that interventions bringing about $X = x_i$ make it less likely to be the case that $Y \neq y_j$, as compared to cases where the intervention does not occur. It is worth noting that there is a substantial debate over whether contrastive explanations can ever be probabilistic, with Lewis (1986) perhaps the most prominent proponent of the view that chancy events cannot be explained contrastively, and Percival (2000) arguing that contrastive explanation is impossible when both fact and foil are chancy events. Though I do not have space here to address this debate in detail, for the sake of argument I am adopting Hitchcock's (1996, 1999) position, *contra* Lewis and Percival, that there is no conflict between contrastive and probabilistic explanation.[5]

Therefore, in order to determine whether one event causally explains another in a probabilistic context, we need to be able to determine whether an intervention

[3] I am leaving to one side issues associated with puzzling cases of pre-emption, overdetermination, and back-ups. For a comprehensive assessment of these cases, see Fenton-Glynn (2017).

[4] See Lipton (1990) for more detail on the fact-foil distinction and its role in contrastive explanation.

[5] Although Hitchcock is, strictly speaking, concerned with the "explanatory relevance" of one event to another, where 'explanatory relevance' is defined so that event a explains event e just in case $p(e|a) \neq p(a)$, I take it that his defense of probabilistic contrastive explanation carries over to an account that conceives of causal explanation as probability-raising under intervention.

bringing about the first event increases the probability of the second event. The Bayes nets formalism provides an explicit methodology for determining whether these conditions for a probabilistic causal explanation are in fact satisfied. To illustrate, let us begin by defining the *Markov factorization* of a Bayes net. Pearl (2000, p. 15-16) proves that if $\mathcal{V} = \{V^1, V^2, ..., V^m\}$, and if each variable in $\mathcal{V}$ has a corresponding value $v_1^1, v_2^2, ..., v_m^m$, and if $\mathcal{PA}_{V^i}$ is the set of parents of a variable V^i in the Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$, and if $\mathbf{pa_{V^i}}$ is the n-tuple of values taken by those parents, and if $\mathcal{N}$ satisfies CMC, then the probability $p(v_1^1, v_2^2, ..., v_m^m)$ can be factorized as follows.

$$p(v_1^1, v_2^2, ..., v_m^m) = \prod_{i=1}^{m} p(v_i^i | \mathbf{pa_{V^i}}) \tag{1.1}$$

Next, suppose that we intervene on a set of variables $\mathcal{X} \in \mathcal{V}$, setting it to the set of values $\mathbf{x_o}$. Pearl (2000, p. 30) and Spirtes et al. (2000, p. 51) show that in a Bayes net that satisfies CMC, the interventional conditional probability $p(v_1^1, v_2^2, ..., v_m^m | do(\mathbf{x_o}))$ can be obtained using the following truncated factorization:

$$p(v_1^1, v_2^2, ..., v_m^m | do(\mathbf{x_o})) = \prod_{i=1}^{m} p_{do(\mathbf{x_o})}(v_i^i | \mathbf{pa_{V^i}}) \tag{1.2}$$

Where each probability $p_{do(\mathbf{x_o})}(v_i^i | \mathbf{pa_{V^i}})$ is defined as follows:

$$p_{do(\mathbf{x_o})}(v_i^i | \mathbf{pa_{V^i}}) = \begin{cases} p(v_i^i | \mathbf{pa_{V^i}}) & \text{if } V^i \notin \mathcal{X} \\ 1 & \text{if } V^i \in \mathcal{X} \text{ and } v_i^i \text{ consistent with } \mathbf{x_o} \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

To illustrate how the do-operator works, consider the Bayes net $Y \leftarrow X \rightarrow Z$, and suppose that we want to calculate the probability $p(z_k | do(y_j))$, where y_j and z_k are any values of Y and Z, respectively. The law of total probability gives us the following, where X has n values:

$$p(z_k | do(y_j)) = \sum_{i=1}^{n} p(z_k, x_i | do(y_j)) \tag{1.4}$$

Applying equation (1.2) yields the following:

$$\sum_{i=1}^{n} p(z_k, x_i | do(y_j)) = \sum_{i=1}^{n} p_{do(y_j)}(z_k | \mathbf{pa_Z}) p_{do(y_j)}(x_i | \mathbf{pa_X}) \tag{1.5}$$

$$\sum_{i=1}^{n} p_{do(y_j)}(z_k | \mathbf{pa_Z}) p_{do(y_j)}(x_i | \mathbf{pa_X}) = \sum_{i=1}^{n} p(z_k | x_i) p(x_i) = p(z_k) \tag{1.6}$$

Thus, the truncated factorization shows us that intervening on Y does not change the probability distribution over Z with respect to the marginal distribution over Z. This is in keeping with the common-cause structure $Y \leftarrow X \rightarrow Z$; interventions on Y do not make a difference to the probability distribution over Z. This is in contrast to the conditional probability $p(z_k|y_j)$, which can be written as follows:

$$p(z_k|y_j) = \sum_{i=1}^{n} p(z_k, x_i|y_j) \tag{1.7}$$

$$\sum_{i=1}^{n} p(z_k, x_i|y_j) = \sum_{i=1}^{n} \frac{p(z_k|x_i)p(x_i|y_j)p(y_j)}{p(y_j)} \tag{1.8}$$

$$\sum_{i=1}^{n} \frac{p(z_k|x_i)p(x_i|y_j)p(y_j)}{p(y_j)} = \sum_{i=1}^{n} p(z_k|x_i)p(x_i|y_j) \tag{1.9}$$

For some possible values of $p(z_k|y_j)$, $p(z_k|y_j) \neq p(z_k)$, so that the observational conditional probability has a different value than the interventional conditional probability.

As a second example, consider the Bayes net $Y \rightarrow X \rightarrow Z$, and suppose that we want to calculate the probability $p(z_k|do(y_j))$. As before, the law of total probability gives the following equation:

$$p(z_k|do(y_j)) = \sum_{i=1}^{n} p(z_k, x_i|do(y_j)) \tag{1.10}$$

Applying equation (1.2) yields the following:

$$\sum_{i=1}^{n} p(z_k, x_i|do(y_j)) = \sum_{i=1}^{n} p_{do(y_j)}(z_k|\mathbf{paz})p_{do(y_j)}(x_i|\mathbf{pa_X}) \tag{1.11}$$

$$\sum_{i=1}^{n} p_{do(y_j)}(z_k|\mathbf{paz})p_{do(y_j)}(x_i|\mathbf{pa_X}) = \sum_{i=1}^{n} p(z_k|x_i)p(x_i|y_j) = p(z_k|y_j) \tag{1.12}$$

Using the truncated factorization equation, interventional conditional probabilities can be calculated from observational probabilities, provided that we know the graphical structure of the Bayes net that the observational joint probability distribution is defined over. Indeed, Huang and Valtorta (2006) show that for any Bayes net that satisfies CMC, interventional conditional probabilities can be derived from the truncated factorization of the joint probability distribution over the Bayes net.

A less formal account of the connection between CMC and interventional conditional probability distributions can be stated as follows. Let $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ be a Bayes net. If we intervene on some variable $X \in \mathcal{V}$, then we make it the case that the value of X no longer depends on its parents, but instead depends solely on the intervention. This can be represented graphically by a sub-graph of $\mathcal{N}$ in which all arrows into X are removed. This sub-graph is called the *pruned sub-graph* of $\mathcal{N}$ for an intervention on X. Spirtes et al. (2000) prove that if $\mathcal{N}$ satisfies CMC, so that we can calculate the joint probability distribution over the pruned sub-graph of $\mathcal{N}$ given an intervention on any variable X, using the equations (1.2) and (1.3). This calculation allows us to determine which types of events represented in a Bayes net explain other types of events represented in the same Bayes net.

1.2.3 Interventions as Events

It is worth noting that the interventional conditional probability $p(y_j|do(x_i))$, where x_i and y_j are any values of the variables in some Bayes net, is not a typical conditional probability by the lights of the Bayes nets framework. It can be calculated via the equation (1.2), but the Bayes nets formalism does not allow it to be manipulated via Bayes theorem to obtain a likelihood or prior probability for the intervention $do(x_i)$. This is because interventions are not, themselves, values of variables in $\mathcal{V}$, and so they are not assigned probabilities by the probability distribution over a Bayes net.

However, we may sometimes wish to speak of the probabilities or likelihoods of interventions themselves. To this end, I propose the following augmentation of the Bayes nets formalism, which follows closely a similar proposal in Pearl (1994) and Pearl (2000, p. 70-72). For any Bayes net $\langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$, all variables in $\mathcal{V}$ are measurable with respect to some probability space $\langle \Omega, \mathcal{A}_\Omega, p(\cdot) \rangle$, and $p(\cdot)$ defines a joint probability distribution over the possible combinations of values for the variables in $\mathcal{V}$. Next, define a set of *intervention variables* $\mathcal{I}^{\mathcal{V}}$ such that $\mathcal{V} \cap \mathcal{I}^{\mathcal{V}} = \emptyset$, where each variable in $\mathcal{V} \cup \mathcal{I}^{\mathcal{V}}$ is measurable with respect to a probability space $\langle \Omega, \mathcal{A}_\Omega, p^*(\cdot) \rangle$, where $p^*(\cdot)$ assigns a joint probability to all combinations of values of all variables in $\mathcal{V} \cup \mathcal{I}^{\mathcal{V}}$, and where $p(\cdot)$ and $p^*(\cdot)$ agree with respect to the probability assigned to all combinations of values of all variables in $\mathcal{V}$. Note that the probability distribution $p^*(\cdot)$ over $\mathcal{V} \cup \mathcal{I}^{\mathcal{V}}$ is not a unique extension of the probability distribution $p(\cdot)$ over $\mathcal{V}$; many such extensions are possible.

For each variable $X \in \mathcal{V}$, there is a variable $I^X \in \mathcal{I}^{\mathcal{V}}$ whose values are just the set of possible interventions on X, plus a value $null$ indicating a lack of an intervention. That is, if the set of possible combinations of values for the variables in X is $\{x_1, x_2, ..., x_n\}$, then the value set of I^X is $\{do(x_1), do(x_2), ..., do(x_n), null\}$. Let us define an extended Bayes net $N^* = \langle \mathcal{V} \cup \mathcal{I}^{\mathcal{V}}, \mathcal{E}^*, p^*(\cdot) \rangle$ where the set of edges $\mathcal{E}^*$ is

such that each $I^X \in \mathcal{I}^{\mathcal{V}}$ has no parents and has as its only child the variable X. For any variable $X \in \mathcal{V}$, let $\mathcal{PA}^*_X$ be the set containing its parents in the extended Bayes net N^*. For any value x_k of X and any n-tuple of values $\mathbf{pa}^*_X$, the following holds:

$$p^*(x_k|\mathbf{pa}^*_X) = \begin{cases} p(x_k|\mathbf{pa}_X) & \text{if } I^X = null \\ 1 & \text{if } I^X = do(x_k) \\ 0 & \text{if } I^X = do(x_l), \text{ where } x_l \neq x_k \end{cases} \tag{1.13}$$

This equation is the sole restriction on the relationship between the probability distributions in the original Bayes net $\mathcal{N}$ and the extended Bayes net $\mathcal{N}^*$. Other features of $\mathcal{N}^*$, e.g. the value of any probability of an intervention $p^*(do(x_k))$, are not determined by the original Bayes net $\mathcal{N}$, and may be chosen arbitrarily or via other, case-specific considerations.

Since all variables in $\mathcal{I}^{\mathcal{V}}$ have no parents, if $V = \{V^1, V^2, \ldots, V^m\}$, then any probability $p^*(v^1_1, v^2_2, \ldots, v^m_m; i^{v1}_1, i^{v2}_2, \ldots, i^{vm}_m)$, where i^{vj} denotes the value taken by a variable I^{Vj} with the value set $\{do(v^j_1), do(v^j_2), \ldots, do(v^j_s), null\}$, can be factorized as follows:

$$p^*(v^1_1, v^2_2, \ldots, v^m_m; i^{v1}_1, i^{v2}_2, \ldots, i^{vm}_m) = \prod_{j=1}^{m} p^*(i^{vj}_j) \prod_{j=1}^{m} p^*(v^j_j|\mathbf{pa_{Vji}}) \tag{1.14}$$

This factorization, along with equation (1.13), allows us to define the probability $p(\mathbf{x_o}|do(\mathbf{y_q}))$, where $\mathbf{x_o}$ is an n-tuple of values taken by the variables in a set $\mathcal{X} \subseteq \mathcal{V}$ and $\mathbf{y_q}$ is an n-tuple of values $\{y^1_1, y^2_2, \ldots, y^s_s\}$ taken by values in the variable set $\mathcal{Y} = \{Y^1, Y^2, \ldots, Y^s\} \subseteq \mathcal{V}$, as follows:

$$\begin{aligned} p(\mathbf{x_o}|do(\mathbf{y_q})) = p^*(\mathbf{x_o}|I^{Y1} = do(y^1_1), I^{Y2} = do(y^2_2), \ldots, I^{Ym} = do(y^m_m); \\ I^{Z1} = null, I^{Z2} = null, \ldots, I^{Zm} = null) \end{aligned} \tag{1.15}$$

Where $\{I^{Z1}, I^{Z2}, \ldots, I^{Zr}\} \subseteq \mathcal{I}^{\mathcal{V}}$ is the set of variables whose values are the possible interventions on variables in the set $\mathcal{V} \setminus \{\mathcal{Y}\} = \{Z^1, Z^2, \ldots, Z^r\}$. The right-hand term of equation (1.15) is an ordinary conditional probability that can be used to assign marginal and posterior probabilities to possible interventions on the Bayes net $\mathcal{N}$. While non-standard, this augmentation of the Bayes nets formalism is arguably tacit in recent work wherein authors assign probabilities to interventions on a Bayes net, e.g. Griffiths et al. (2015), Pocheville et al. (2017), and Eva and Stern (forthcoming).

This completes my exegesis of the Bayes nets approach to causal representation, and how it facilitates an interventionist approach to probabilistic causal explanation. In

what follows, I show that this approach is consistent with an ecumenist response to various causal exclusion arguments.

1.3 The Case for Ecumenism in Bayes Nets

Causal ecumenism is the claim that the same event can be causally explained at multiple levels of granularity. This claim can be made more precise, in order to be substantively defended. Let X be any random variable. Recall that X is a *coarsening* of some other random variable X' just in case X is not identical to X', and the value set of X is a *quotient set* of the value set of X'. One set A is a quotient set of another set B just in case A contains all and only the equivalence classes of B according to some equivalence relation defined over B. Throughout this chapter, if one variable is a coarsening of another, then the more fine-grained variable will be indicated via a superscript $\prime$. The same holds for graphs containing more fine-grained variables, and probability distributions over more fine-grained variables. If X is a coarsening of X', then X' is a *refinement* of X. We can now state the thesis of causal ecumenism in the Bayes nets context as follows.

> **CE:** Let it be the case that $C' = c'_l$ causally explains why $E' = e'_s$ according to a Bayes net $\mathcal{N}' = \langle \mathcal{V}', \mathcal{E}', p(\cdot) \rangle$, where $\mathcal{N}'$ is an adequate representation of some target system Φ, and where $\mathcal{N}'$ satisfies CMC and Minimality, and where $E' \in \mathcal{V}'$ and $C' \in \mathcal{V}'$. In some cases in which these preliminary conditions are satisfied, there is a Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ that is also representationally adequate with respect to Φ such that: 1) $\mathcal{V} = \{C\} \cup \mathcal{V}' \setminus \{C'\}$, 2) the joint probability distribution $p(\cdot)$ over the variables in $\mathcal{V} \setminus \{C'\}$ is identical in both Bayes nets, 3) $\mathcal{N}$ satisfies CMC and Minimality, 4) C is a coarsening of C', 5) if $C' = c'_l$ implies that $C = c_j$, then $C = c_j$ causally explains why $E' = e'_s$.

Thus, under causal ecumenism, if we can give a causal explanation of some event at one level of granularity, then it is sometimes possible that we can give a causal explanation of the same event at a coarser level of granularity, by replacing the explanatory variable with a coarsening of itself. Given that the causal exclusion arguments considered here claim that there is always one correct level of causal explanation for a given explanandum, the possibility of causal explanation at multiple levels of explanation is sufficient to establish the truth of causal ecumenism.

As causal ecumenism is an existentially quantified thesis, its truth can be demonstrated by an example. Suppose that the graph in Figure 1.4 is an adequate representation of the causal structure of the weather system in a given region. A' is a variable

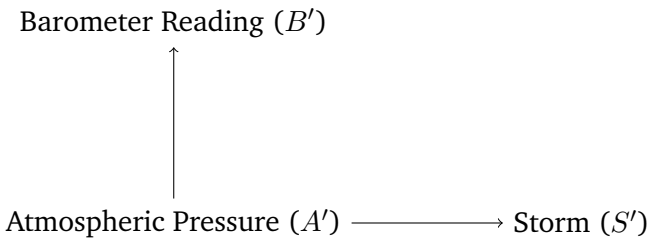

Fig. 1.4: Weather System Graph with Fine-Grained Pressure Variable

$p(A' = low, B' = high, S' = yes) = .042$
$p(A' = low, B' = high, S' = no) = .003$
$p(A' = low, B' = low, S' = yes) = .267$
$p(A' = low, B' = low, S' = no) = .021$
$p(A' = medium, B' = high, S' = yes) = .1$
$p(A' = medium, B' = high, S' = no) = .1$
$p(A' = medium, B' = low, S' = yes) = .067$
$p(A' = medium, B' = low, S' = no) = .067$
$p(A' = high, B' = high, S' = yes) = .042$
$p(A' = high, B' = high, S' = no) = .267$
$p(A' = high, B' = low, S' = yes) = .003$
$p(A' = high, B' = low, S' = no) = .021$

Tab. 1.1: Joint Distribution for Fine-Grained Weather System

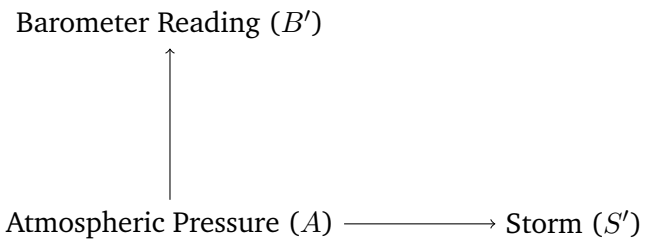

Fig. 1.5: Weather System Graph with Coarse-Grained Pressure Variable

representing atmospheric pressure, with possible values $\{low, medium, high\}$. B' is a variable representing the reading of a barometer, with possible values $\{low, high\}$. S' is a variable representing whether or not a storm occurs, with possible values $\{yes, no\}$. The joint probability distribution over these variables is given in Table 1.1. Examining this joint distribution, one can see that low pressures, low barometer readings, and the presence of storms are correlated, as are high pressures, high barometer readings, and the absence of storms. However, there is some measurement error, such that there is not a perfectly deterministic relationship between atmospheric pressure and readings. Similarly, there is some stochasticity in the relationship between atmospheric pressure and the presence or absence of storms. One can check that this distribution satisfies CMC with respect to the graph shown in 1.4. Thus, we are able to calculate the conditional probability distribution over variables in the graph, given interventions on other variables.

Suppose that $A' = low$ and $S' = yes$. The following calculation shows that the fact that atmospheric pressure is low causally explains why a storm occurred (details of this calculation, and all subsequent calculations, are given in the appendix):

$$p(S' = yes|do(A' = low)) = .93 > p(S' = yes) = .52 \tag{1.16}$$

By contrast, the following calculation shows that if the barometer reading is such that $B' = low$, then this does not explain why $S' = yes$:

$$p(S' = yes|do(B' = low)) = .52 = p(S' = yes) = .52 \tag{1.17}$$

Thus, this case is consonant with a commonsense understanding of meteorology; while low atmospheric pressure can causally explain why a storm occurs, a low reading on a barometer cannot, in itself, causally explain why a storm occurs.

Now suppose that we coarsen the variable A'. To do this, we define an equivalence relation over the value set $\{low, medium, high\}$ such that 'low' and 'medium' are equivalent to each other as well as themselves, and 'high' is equivalent only to itself. Defining a quotient set of the fine-grained value set according to this equivalence re-

$p(A = low/medium, B' = high, S' = yes) = .142$
$p(A = low/medium, B' = high, S' = no) = .103$
$p(A = low/medium, B' = low, S' = yes) = .333$
$p(A = low/medium, B' = low, S' = no) = .088$
$p(A = high, B' = high, S' = yes) = .042$
$p(A = high, B' = high, S' = no) = .267$
$p(A = high, B' = low, S' = yes) = .003$
$p(A = high, B' = low, S' = no) = .021$

Tab. 1.2: Joint Distribution for Coarse-Grained Weather System

lation results in a coarse-grained variable A with the value set $\{low/medium, high\}$. If we replace the variable A' in Figure 1.4 with the variable A, we get the graph in Figure 1.5 and the joint distribution over that graph shown in Table 1.2. This joint distribution assumes that, in an extended Bayes net that includes intervention variables, as described in Section 1.2.3, $p^*(A' = low|I^A = do(A = low/medium)) = p^*(A' = medium|I^A = do(A = low/medium))$; this assumption is made for mathematical tractability and is not generally required for the result stated in Proposition 3. This joint distribution is calculated under the stipulation that if $\{x'_1, x'_2, ..., x'_k\}$ are all and only those values of a fine-grained variable that stand in some equivalence relation to each other and are all mapped to the coarse-grained variable value x_r, then $p(x_r) = \sum_{i=1}^{k} p(x'_i)$. One can check that the joint distribution in Table 1.2 satisfies CMC with respect to the graph in Figure 1.5. Thus we can confirm that if $S' = yes$, then this can be explained by the fact that $A = low/medium$, by calculating the following:

$$p(S' = yes|do(A = low/medium)) = .631 > p(S' = yes) = .52 \tag{1.18}$$

This calculation shows that causal ecumenism is satisfied in this case. The proposition that $A' = low$ causally explains why $S' = yes$ according to the Bayes net described by the graph in Figure 1.4 and the joint distribution in Table 1.1. The Bayes net described by the graph in Figure 1.5 and the joint distribution in Table 1.2 is an adequate Bayes net representation of the same weather system such that: 1) the only difference in variable sets between the two graphs is the replacement of A' with its coarsening A, 2) the joint probability distribution over S' and B' is the same in both Bayes nets, 3) the Bayes net satisfies CMC and Minimality, 4) A is a coarsening of S', and 5) $A' = low$ implies that $A = low/medium$ and $A = low/medium$ causally explains why $S' = yes$. Thus, all the conditions for causal ecumenism are satisfied.

There is a more general result that speaks in favor of causal ecumenism. Specifically, the following proposition is true (see appendix for proof):

Proposition 3. *For any Bayes net $\mathcal{N}' = \langle \mathcal{V}', \mathcal{E}', p(\cdot)\rangle$ that satisfies CMC and Minimality, and any variable $C' \in \mathcal{V}'$ such that C' has more than two values in its value set, if $C' = c'_l$ causally explains why $E' = e'_s$, where $E' \in \mathcal{V}'$, then there is a Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot)\rangle$ that satisfies the five conditions listed in **CE**.*

This proposition shows that any variable with more than two values can be coarsened into a variable with fewer values, while preserving any explanatory relations between the value taken by the coarsened variable and the value taken by other variables in the graph. If, in this coarsening process, the explanatory value c'_l is deemed to be equivalent to some other fine-grained value c'_u, then replacing $C' = c'_l$ with the more coarse-grained explanans $C = c_j$ will result in a more abstract explanation than its fine-grained counterpart, in Weslake's (2010) sense of having more possible instantiations in various target systems that the model might be used to represent. This occurs in the case presented above; $A' = low$ and $A' = medium$ were deemed equivalent, and the explanation that resulted from grouping them together in a coarsening ($S' = yes$ because $A = medium/low$) was more abstract. Additionally, the existence of a more coarse-grained version of some fine-grained causal explanations implies that the reverse is also true; for some coarse-grained variables, one can find a way of refining the values of that variable so as to achieve a more fine-grained causal explanation (providing that some hypothetical limit to the fine-grainedness of the description of any system has not been reached).

To clarify, Proposition 3 does not imply that any causal explanation can be restated in more coarse-grained, abstract language. Rather, it says only that variables with more than two values can be replaced by their coarsenings while maintaining our ability to explain why other variables in the graph take the values that they do. To illustrate this distinction, suppose that in some Bayes net presenting whether or not flooding will occur in a region, there is a three-valued variable R' representing the amount of rainfall in a region, with possible values $\{low, medium, high\}$, and two-valued variable F' representing whether a flood occurs, with possible values $\{no, yes\}$. Suppose further that $R' = high$ explains why $F' = yes$. Under some joint probability distributions, it may be the case that if we coarsen R' into R so that it can take the values $\{low, medium/high\}$, then it is not the case that $R = medium/high$ explains why $F' = yes$. Similarly, if we coarsen R' into a variable R that can take the values $\{low/high, medium\}$, then it is not the case that $R = low/high$ explains why $F' = yes$. Having ruled out these options, Proposition 3 does imply that we can coarsen R' into a variable R that can take the values $\{low/medium, high\}$, such that $R = high$ explains why $F' = yes$. This leaves us with effectively the same explanation as we had pre-coarsening, albeit one that is embedded within a more coarse-grained model of the target system as a whole.

Additionally, Proposition 3 does not imply that any coarsening of the causal variable will preserve the explanatory relationship between values taken by the causal variable and the values taken by its effect variable(s). Indeed, some poorly-chosen coarsenings will result in a situation such that interventions on a fine-grained variable X' change the probability distribution over another variable Y', but when X' is replaced by a coarsening X, no interventions on X change the probability distribution over Y'. Consider the following case, based on an example due to Spirtes and Scheines (2004). In humans, the presence of high-density lipids in the bloodstream decreases the risk of heart attack, whereas the presence of low-density lipids in the bloodstream increases the risk of heart attack. Let L' be a variable whose value set contains pairs of real numbers denoting the amount of each lipid (in $\mathrm{mg/dL}$) in a patient's bloodstream. So if a patient's blood contains 100 $\mathrm{mg/dL}$ of low-density lipids and 160 $\mathrm{mg/dL}$ of high-density lipids, then $L' = \langle 100, 160 \rangle$. Let H' be a variable representing whether or not someone has a heart attack. Clearly, there are some interventions on L' such that conditioning on those interventions increases the probability that $H' = yes$ or $H' = no$. However, suppose that we replace L' with its coarsening L, where the value of L is just the sum of the patient's two lipid levels (e.g. if $L' = \langle 100, 160 \rangle$, then $L = 260$). There are marginal probability distributions over the fine-grained variable L' such that, for any intervention on L, conditioning on that intervention does not raise or lower the probability of heart attack. This is because the total amount of lipids in the patient's bloodstream is not informative as to the patient's risk of heart attack if we do not also know the density of those lipids. These sorts of cases show that in choosing coarse-grained variables, scientists have to be careful to choose coarsenings that still allow for the possibility of causal explanation in the more coarse-grained model. Proposition 3 shows that this can be done in principle, as long as the coarsening is defined in terms of an appropriate equivalence relation between values of the fine-grained variable.

However, this example does bring to the foreground an important feature of causal explanation and causal ecumenism. Suppose that, in general, interventions that set a person's overall lipid count to high levels tend to do so by increasing the amount of low-density lipids in that person's bloodstream, while leaving fixed the amount of high-density lipids. Thus, an intervention setting L to a high value, e.g. $L = 300$, would be evidence of a person being more likely to have a high level of low-density lipids. This would make it the case that $p(H' = yes|do(L = 300)) > p(H' = yes)$. If we assume that L and H' are variables in a Bayes net that satisfies CMC and Minimality, then we get the result that if someone has a heart attack and their total lipid level is 300 $\mathrm{mg/dL}$, then their total lipid level explains why they had a heart attack.

At first glance, this seems like a strange result. After all, interventions on the total lipid count only raise a person's likelihood of having a heart attack in virtue of

the fact that these interventions tend to also increase the amount of low-density lipids in their bloodstream. So really, it seems that it is this fact about low-density lipids, rather than the bare fact of the person's overall lipid count, that explains their heart attack. However, according to the ecumenism that I have defended above, the correct analysis of this case is as follows. It is true that, on a more fine-grained model, it is really the person's high level of low-density lipids that explains their heart attack. However, it is also the case that, on a more coarse-grained model, their total lipid count explains their heart attack. To say that both models can be explanatory is just to accept the thesis of causal ecumenism. However, this does not entail that the two models furnish *equally good* explanations of the person's heart attack. Indeed, I would hold that in this case, the more fine-grained explanation is the better one. It is just that both putative explanations do satisfy the conditions to be an explanation of the person's heart attack.

Biting this bullet in these cases allows us to give the same response in cases of scientific explanation at different levels of granularity where such a response appears to be warranted. Suppose that we heat a box of gas, and thereby cause an increase in the pressure that the gas exerts on the box. We can explain the increased pressure on the box by citing the increase in heat brought about via an intervention. At a more fine-grained level, we can say that the increased average kinetic energy of the particles of gas in the box explains the pressure that the gas exerts on the box. However, both of these explanations only make sense under the assumption that the increase in average kinetic energy of the particles in the box is at least somewhat evenly distributed between the particles. It may be that, as a result of heating, the particles in the center of the box are very high-energy, and the particles further from the center of the box maintain their kinetic energy from before the box is heated. Average kinetic energy increases, but is not evenly distributed. Under these conditions, the pressure on the box would not increase, since particles near the edge of the box would not be colliding into it with greater velocity.

It is scientifically respectable to explain the increase in the box's temperature by citing its increased heat; we do not need to specify that the box is heated in a way such that the particles furthest from the center of the box increase their kinetic energy. This is because heating a box tends to result in a roughly uniform increase in the average kinetic energy of particles in all regions of the box, so that the scenario described above in which only the particles in the center of the box have increased energy tends not to occur. This matches the state of affairs in the second lipid case above. By stipulation, increases in total lipid count due to an increase in high-density lipids tend not to occur, just as increases in heat attributable solely to an increased kinetic energy of molecules close to the center of a box of gas also tend not to occur. When conditionalizing on a coarse-grained intervention changing a person's total lipid count, or changing the heat of a box of gas, we can assume that, at the fine-grained

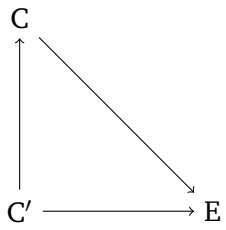

Fig. 1.6: Multi-Level Causal Graph

level, the intervention is realized in a manner that is consistent with an empircally supported conditional probability distribution over the fine-grained variable, given interventions on its coarsening. Thus, in both cases, the coarse-grained explanation is justified, even if this justification relies on the coarse-grained explanans having its more typical fine-grained instantiation.

I take these examples and results to motivate causal ecumenism as a philosophically and scientifically respectable thesis that is in keeping with both the Bayes nets formalism and the interventionist approach to causal explanation. In what follows, I consider and respond to two ways of motivating the rival thesis that there is a uniquely correct level of causal explanation for a given event.

1.4 Against Upward Exclusion

Gebharter's (2017) argument for upward exclusion takes its cues from Kim, but uses the formal machinery of Bayes nets. His argument begins with the following stipulations: let E be some effect variable, and let C and C' be two possible causal variables such that C is a coaresening of C'. Gebharter posits that we can represent all three variables as part of the same graph, as shown in Figure 1.6. This stands in contrast to my account of coarse-graining, which represents different granularities of explanation using different graphs. Gebharter points out that the graph in Figure 1.6 violates Minimality. To see why, note that even though there is an edge from C to E, the value of E is independent of the value of C, provided that we conditionalize on C'. This is because the probability distribution over both C and E will be totally determined by the value of C'. We can remove the edge from C to E, creating a proper sub-graph of the graph in Figure 1.6 that does not violate CMC. Thus, the graph in Figure 1.6 violates Minimality. By contrast, the proper subgraph induced by removing the edge from C to E does satisfy Minimality. Via this reasoning, Gebharter concludes that the causal relationship between C' and E excludes the possibility of a causal relationship between C and E, according the the Bayes nets axioms.

To illustrate the upshot of Gebharter's conclusion, suppose that we throw a ball at a window and it breaks. Suppose further that we know the microphysical properties of the ball as it makes impact with the window. These microphysical properties allow us to assign a probability distribution over whether the window breaks. These microphysical properties also fix the macrophysical properties of the ball, such as its velocity and mass. So if we conditionalize on the microphysical properties of the ball, we see that whether or not the window breaks does not depend on the macrophysical properties of the ball. Thus, Gebharter would conclude, we cannot simultaneously explain the broken window by citing both the microphysical and macrophysical properties of the ball.

The premise that is doing the bulk of the work here is the claim that coarse-grained events and their fine-grained realizers should be represented using the same graph, and that fine-grained realizers should be represented as causing the coarse-grained properties that they realize. Gebharter's argument to this effect proceeds as follows. He begins by pointing out, correctly, that because Bayes nets satisfy CMC, the causal relations within a given Bayes Net have the property of *stability*. By stability, Gebharter means that for some causal relation $C \rightarrow E$, where C is the only parent of E, the conditional probability distribution over C, given E, is independent of the prior probability distribution over E. For example, the probability that a ball is thrown at a window, given that the window breaks, does not depend on the prior probability that the window breaks. Gebharter argues that the supervenience relations that hold between coarse-grained events and their fine-grained realizers also exhibit this stability property. For example, the conditional probability that a ball has some microphysical properties, given its velocity, is not effected by the prior probability that the ball has its particular velocity. Since this stability property seems to be an essential feature of both supervenience relations and causal relations, Gebharter concludes that supervenience relations and causal relations should have the same formal representation in a Bayesian Network.

Up until his conclusion, everything that Gebharter says is correct. However, for his argument to be valid, it would have to be the case that sharing this stability property implies that two relations are indistinct. I argue that this assumption is false. Even though causation and supervenience relations both exhibit this stability property, there are other properties that distinguish the two types of relations. First, causal relations necessarily exhibit an asymmetry of probabilistic dependence under interventions. For example, intervening on the velocity of the ball thrown at a window changes the probability that the window breaks, but intervening so that a window does not break does not change the velocity of the ball thrown at it. This same asymmetry is not necessarily a property of supervenience relations; an intervention that changes the microphysical properties of the ball can also change

its velocity, and a change in the velocity of the ball must result in a change in its microphysical realization.

Second, there are metaphysical reasons for thinking that supervenience relations are fundamentally different from causal relations. To begin, token causal relations hold between *distinct* events, whereas token supervenience relations only hold between properties of objects or events that are in some sense not wholly distinct. When I say that the solidity of my desk supervenes on its molecular structure, I am describing a relation between two properties of the same token object. It would not make sense to say that the molecular structure of my desk *causes* it to be solid. Similarly, if I throw a brick at a window and the window breaks, it would be natural to say that my throwing the brick caused the window to break, but wrong to say that the window breaking supervenes on my throwing the brick. Also, there is a difference regarding the role that temporal ordering plays in characterizing the two relations; even if we do not want to rule out simultaneous or backwards causation, it is certainly the case that many causes precede their effects temporally, whereas subvening and supervening events necessarily occur simultaneously.

Gebharter disagrees with this line of argument, asserting that one could "describe the temperature of a gas in a tank as the effect of the behavior of the gas particles in the tank" (2017, 8). However, this assertion is presented without argument. While I agree that we might speak this way about temperature and gas particles, it could nevertheless be strictly incorrect to do so, for the reasons given above. Although Gebharter is correct that supervenience and causation are both stable in the way that he describes, there remain important differences between the two relations (e.g. differences regarding symmetry or asymmetry of dependence under interventions on either relatum) that clearly matter in a Bayes nets context. One might try to defend Gebharter here on the basis that supervenience and other constitutive relations are essentially "metaphysical" versions of causation, in contrast with "physical" causation, which holds between distinct events; a version of this view is defended by both Schaffer (2016) and Wilson (2018). My response here is just to say that even if supervenience is a metaphysical variant of causation, this does not compel us to represent metaphysical and physical causal relations on the same graph. Thus, Gebharter's key premise rests on a confusion about how to represent the relation between variables in a Bayes net and coarsenings of those same variables. Gebharter assumes that they should be represented as variables in the same Bayes net, while I stipulate that they must be represented as variables in different Bayes nets. For the reasons given above, I take it that my position is the better-motivated one.

1.5 Against Downward Exclusion

List and Menzies (2009) propose their own difference-making semantics for causation. In some cases where I have argued that ecumenism prevails, List and Menzies' account would entail that there is downward exclusion. To defend my ecumenical approach, I will have to say where I believe that their account goes wrong. In this section, I will argue that their defense of downward exclusion relies on an account of causal explanation that is too demanding to account for all of the cases that we care about.

List and Menzies hold that some event $C = c_j$ makes a difference to an event $E = e_i$ just in case the following two conditions hold: first, if $C = c_j$ "in any relevantly similar possible situation", then $E = e_i$; second, if $C \neq c_j$ in any relevantly similar possible situation, then $E \neq e_i$ (List and Menzies 2009, 482). In laying out their approach, List and Menzies only consider cases in which the causal relationships between variables are deterministic, but their account can be straightforwardly extended to the probabilistic setting. To do this, we can just say that the event $C = c_j$ makes a difference to the event $E = e_i$ just in case the probability of $E = e_i$ is greater when $C = c_j$ than it is when C takes any of its other possible values. When this difference-making relationship holds between all values of C and E, the two variables are said to be *proportional* to one another.[6] Additionally, List and Menzies make it clear that they mean for their account to be consistent with the interventionist understanding of causation (2009, p. 481). In what follows, I assume that the values of all causal variables are set via interventions.

Following Yablo (1992), List and Menzies hold that difference-making and proportionality are necessary conditions for causal explanation. To illustrate this commitment, they put forward the following example:

> Suppose, for example, there is a drug that causes patients to recover from an illness. The effect variable is a binary variable whose values are recovery or non-recovery. But the cause variable is many-valued, with possible values 0mg, 50mg, 100mg, 150mg, and 200mg. Suppose that any regular dose at or above 150mg cures a patient, but any lower dose does not. Suppose a patient has taken a regular dose of 150mg and has recovered from the illness. What made the difference to the

[6] This definition of proportionality in the probabilistic context is found in Pocheville et al. (2017). While one may worry that this definition is too weak, I do not consider here possible alternative definitions. I also assume for the sake of argument that there are not infinitely many values of C that confer increasingly high conditional probabilities to some value of E, in which case this definition would counter-intuitively imply that no values of C are difference-makers with respect to any values of E.

> patient's recovery? [...] The answer is 'Giving the patient a dose of at least 150mg'. Other answers are either too specific, or not specific enough. For example, the cause cannot be 'Giving the patient a dose above 50mg' because [...] some relevantly similar patients who are given a dose above 50mg, say 100mg, do not recover. Similarly, it cannot be 'Giving the patient a dose of exactly 150mg' because [...] some relevantly similar patients who are not given a dose of exactly 150mg—say they are given 200mg—nonetheless recover (2009, pp. 482-3).

The upshot of this example is as follows. List and Menzies hold that every change in the causal variable has to make a difference to the probability that the effect variable takes its actual value, or else there is not difference-making, and therefore there is not causal explanation. It does not make sense, on their view, to say that giving the patient *exactly* 150mg of the drug caused their recovery, because giving them some other amount above 150mg would have made no difference to their probability of survival. To capture the difference-making features of the explanans variable, we need to say only that the patient was given *at least* 150mg of the drug. So the only acceptable explanation of why the patient recovered is that they were given at least 150mg of the drug. Put another way, List and Menzies hold that the only acceptable variable for use in a causal model of the patient's recovery is a two-valued variable specifying whether or not the patient is given at least 150mg of the drug.

This is a putative case of downward causal exclusion. The coarse-grained explanation 'the patient was given at least 150mg of the drug' excludes the more fine-grained explanation 'the patient was given exactly 150mg of the drug.' In these sorts of cases, i.e. cases where higher-level causal explanations are not sensitive to some changes in their lower-level realization, List and Menzies hold that causal explanation only occurs at the higher level. By contrast, the ecumenical position that I have laid out above would hold that both explanations are acceptable. This is because both explanations cite a causal factor—either setting the dosage to at least 150mg or setting the dosage to exactly 150mg—that makes a difference to the patient's probability of recovery. It does not matter, on my view, that some other interventions on this many-valued variable do not make any difference to the patient's probability of recovery. It is worth noting that nothing about List and Menzies' proportionality constraint is entailed by the Bayes nets formalism, which only provides the formal apparatus necessary to calculate intervention distributions over variables in a Bayes net, given interventions on other variables. Thus, in putting proportionality forward as a necessary condition on causal explanation, List and Menzies are proposing an additional axiom for causal reasoning, over and above those described in Section 1.2.

My view is that the difference-making condition that List and Menzies place on causal explanation is too strong. Consider the following example, based on a similar example by Shapiro and Sober (2012). Suppose that the value taken by a variable Y is determined by a parabolic function of the value taken by another variable X, where this function is defined by the equation $Y = 10X - X^2$. Suppose further that it happens to be the case that $X = 2$ and $Y = 16$. If we want to explain why $Y = 16$, it seems natural to say that Y takes this value because $X = 2$. However, this explanation will not satisfy List and Menzies' difference-making desideratum. If $X = 8$, then $Y = 16$. So there are at least some changes in the value of the explanans variable that make no difference to the explanandum variable. It is strange to say here that when $X = 2$, this fact alone does not explain why $Y = 16$. List and Menzies might say that the correct explanation here is that $Y = 16$ because it is the case that either $X = 2$ or $X = 8$, but, as Franklin-Hall (2016) notes, this move seems divorced from scientific practice, where we generally do not see disjunctive explanations, especially when one disjunct is known not to hold.[7]

Although Shaprio and Sober do not consider probabilistic cases, I argue that their upshot still applies in the probabilistic context. Suppose that a person who develops lung cancer smoked a particular brand of cigarette for many years. Suppose further that this brand of cigarette is no more or less carcinogenic than other brands. It would obviously be correct to say that the fact that the person smoked causally explains why they developed lung cancer, since smoking cigarettes causes an increase in the probability of lung cancer in general. However, it is also the case that there is nothing defective, from a purely explanatory perspective, with saying that the fact that the person smoked this particular brand explains why they developed lung cancer. The additional detail about branding may be unnecessary, but it is still true that smoking the specific brand of cigarettes caused the person to develop lung cancer; after all, this is precisely what occurred in the case that I have described. However, if we follow List and Menzies' dictum that causal explanations must include only those features of the explanans that make a difference to the explanandum, then we would have to rule out as a causal explanation any explanation of the person's lung cancer that cites the brand that they smoked. This strikes me as a highly onerous constraint on causal explanation that is not in keeping with scientific or common-sense practice. Thus, I conclude that List and Menzies are only able to widen the class of cases in which downward exclusion occurs by adopting an unintuitive and undesirable account of causation.[8]

[7] In keeping with the ecumenical line developed above, I would hold that the disjunctive explanation is acceptable, though highly non-standard, in this case. However, *contra* List and Menzies, I maintain that this disjunctive explanation does not exclude the finer-grained explanation which says only that $Y = 16$ because $X = 2$.

[8] To the extent that List and Menzies derive their conclusion from counterfactual accounts of causation due to Lewis (1986), these examples are problems for Lewis' view as well.

McDonald (ms) argues in favor of biting the bullet here and insisting that '$X = 2$ or 8' is the true causal explanation of why $Y = 16$ in this context. This move is motivated by a Gricean theory of communication, according to which a contributor to a conversation must "make [their] contribution as informative as is required (for the current purposes of exchange)...[and no] more informative than is required" (Grice 1989, pp. 26-27). That is, McDonald claims that non-proportional causal explanations fail because they specify more information (i.e. the specific value of X) than is needed for a given explanation. My response is to say that even if we grant that satisfying this Gricean maxim of communication is a necessary condition for an *ideal* causal explanation, this is not the same thing as being a necessary condition for any causal explanation at all. Just as we can communicate somewhat successfully even if we are not always compliant with the Gricean maxim above, we can causally explain phenomena in science even if we are not compliant with some notions of what a causal explanation ought to be. The explanation may be sub-optimal, but I hold that it is an explanation nevertheless. Thus, the proportionality constraint on causal explanations cannot be used to motivate a causal exclusion argument.

1.6 Conclusion

I have argued that the Bayes Nets approach to causal explanation is generally consistent with an ecumenical approach to the relationship between granularity and causal explanation. That is, there are many cases in which the same event can be described with varying degrees of granularity, such that both descriptions are explanatory. While this ecumenism does not hold in all cases, it holds in cases where the coarse-graining of the causal variable is done judiciously, so as to preserve causal explanation. This ecumenical view stands opposed to both List and Menzies' view on the one hand, and Gebharter's view on the other. I have argued that both of these views face serious issues, such that ecumenism is preferable.

1.7 Appendix

1.7.1 Calculations

Calculation of Inequality (1.16)

For the purpose of these calculations, *yes*, *no*, *low*, *medium*, and *high* are shortened to y, n, l, m and h, respectively. We begin with left-hand term of the inequality in (1.16). From (1.2), we have:

$$p(S' = y|do(A' = l)) = p_{do(A'=l)}(S' = y|A' = l) \tag{1.19}$$

Which can be calculated as follows:

$$p(S' = y|do(A' = l)) = \frac{p(A' = l, S' = y)}{p(A' = l)} \tag{1.20}$$

$$p(S' = y|do(A' = l)) = \frac{p(A' = l, B' = h, S' = y) + p(A' = l, B' = l, S' = y)}{p(A' = l)} \tag{1.21}$$

$$p(S' = y|do(A' = l)) = \frac{.042 + .267}{p(A' = l)} \tag{1.22}$$

We can calculate $p(A' = l)$ as follows:

$$\begin{aligned} p(A' = l) = {} & p(A' = l, B' = h, S' = y) + p(A' = l, B' = h, S' = n) \\ & + \ p(A' = l, B' = l, S' = y) + p(A' = l, B' = l, S' = n) \end{aligned} \tag{1.23}$$

$$p(A' = l) = .042 + .003 + .267 + .021 = .333 \tag{1.24}$$

This yields the result:

$$p(S' = y|do(A' = l)) = \frac{.042 + .267}{.333} \approx .93 \tag{1.25}$$

Next, we calculate the right-hand term:

$$\begin{aligned} p(S' = y) = {} & p(A' = h, B' = h, S' = y) + p(A' = h, B' = l, S' = y) \\ & + p(A' = m, B' = h, S' = y) + p(A' = m, B' = l, S' = y) \\ & + p(A' = l, B' = h, S' = y) + p(A' = l, B' = l, S' = y) \end{aligned} \tag{1.26}$$

$$p(S' = y) = .042 + .003 + +.1 + .067 + .042 + .267 = .52 \tag{1.27}$$

Calculation of Equation (1.17)

As we have already calculated the right-hand term, we focus on the left-hand term. From the law of total probability, we have:

$$p(S' = y|do(B' = l)) = p(S' = y, A' = l|do(B' = l)) + p(S' = y, A' = m|do(B' = l)) + p(S' = y, A' = h|do(B' = l)) \quad (1.28)$$

From (1.2) and (1.3), we have:

$$p(S' = y|do(B' = l)) = p_{do(B'=l)}(S' = y|A' = l)p_{do(B'=l)}(A' = l) \; p_{do(B'=l)}(S' = y|A' = m)p_{do(B'=l)}(A' = m) + p_{do(B'=l)}(S' = y|A' = h)p_{do(B'=l)}(A' = h) \quad (1.29)$$

$$p(S' = y|do(B' = l)) = p(A' = l, S' = y) + p(A' = m, S' = y) + p(A' = h, S' = y) \quad (1.30)$$

$$\begin{aligned} p(S' = y|do(B' = l)) = {} & p(A' = l, B' = l, S' = y) + p(A' = l, B' = h, S' = y) \\ & + p(A' = m, B' = l, S' = y) + p(A' = m, B' = h, S' = y) \\ & + p(A' = h, B' = l, S' = y) + p(A' = h, B' = h, S' = y) \end{aligned} \quad (1.31)$$

$$p(S' = y|do(B' = l)) = p(S' = y) \approx .52 \quad (1.32)$$

Calculation of Inequality (1.18)

We begin with left-hand term of the inequality in (1.18). From (1.2), we have:

$$p(S' = y|do(A' = l/m)) = p_{do(A'=l/m)}(S' = y|A' = l/m) \quad (1.33)$$

Which can be calculated as follows:

$$p(S' = y|do(A' = l/m)) = \frac{p(A' = l/m, S' = y)}{p(A' = l/m)} \quad (1.34)$$

$$p(S' = y|do(A' = l/m)) = \frac{p(A' = l/m, B' = h, S' = y) + p(A' = l/m, B' = l, S' = y)}{p(A' = l/m)}$$

(1.35)

$$p(S' = y|do(A' = l/m)) = \frac{.142 + .333}{p(A' = l)} \tag{1.36}$$

We can calculate $p(A' = l)$ as follows:

$$\begin{aligned} p(A' = l/m) = p(A' = l/m, B' = h, S' = y) + p(A' = l/m, B' = h, S' = n) \\ + \ p(A' = l/m, B' = l, S' = y) + p(A' = l/m, B' = l, S' = n) \end{aligned} \tag{1.37}$$

$$p(A' = l/m) = .142 + .103 + .333 + .088 = .666 \tag{1.38}$$

This yields the result:

$$p(S' = y|do(A' = l)) = \frac{.142 + .333}{.666} = .642 \tag{1.39}$$

Next, we calculate the right-hand term:

$$\begin{aligned} p(S' = y) = p(A' = h, B' = h, S' = y) + p(A' = h, B' = l, S' = y) \\ + p(A' = l/m, B' = h, S' = y) + p(A' = l/m, B' = l, S' = y) \end{aligned} \tag{1.40}$$

$$p(S' = y) = .042 + .003 + .142 + .333 = .52 \tag{1.41}$$

1.7.2 Proof of Proposition 1

Proof. First, we show that CMC implies Screening Off. CMC states that for any variables X and Y in $\mathcal{N}$, if Y is a non-descendant of X, then X and Y are independent, conditional on $\mathcal{PA}_X$. Let X and W be two variables in $\mathcal{N}$ such that W is a parent of some variable in $\mathcal{PA}_X$. Since W is a non-descendant of X, CMC entails that W is independent of X, conditional on $\mathcal{PA}_X$. Thus, for any two variables X and W in $\mathcal{N}$, if W is a parent of some variable in $\mathcal{PA}_X$, then W is independent of X, conditional on $\mathcal{PA}_X$.

Next, we prove that CMC implies Common Cause. For this, it is helpful to introduce the notion of *d-separation*. To do this, we must first introduce the notion of a *path* and a *collider*. Consider the set of undirected edges (i.e. unordered pairs) $\mathcal{E}^*$ that is consistent with $\mathcal{E}$. There is a path between two variables X and Y if and only if they are connected by a series of undirected edges in $\mathcal{E}^*$. A variable Z is a collider

along the path between X and Y if and only if Z is a child of at least two variables along the path, according to the directed ordering $\mathcal{E}$. A path between two variables X and Y is d-separated by a (possibly empty) set of variables $\mathcal{Z}$ if and only if: i) the path between X and Y contains a non-collider that is in $\mathcal{Z}$, or the path contains a collider, and neither the collider nor any descendant of the collider is in $\mathcal{Z}$.

Next, we introduce the following lemma, which is proved by Verma and Pearl (1990b).

Lemma 1. *A graph $\mathcal{N}$ satisfies CMC if and only if, for any variable set $\mathcal{Z}$ that d-separates any pair of variables X and Y along all paths between them, X and Y are independent conditional on $\mathcal{Z}$.*

We are now in a position to show that CMC implies Common Cause. Let $\mathcal{X}$ and $\mathcal{Y}$ be disjoint sets of variables in a graph $\mathcal{N}$ such that for any $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, X and Y are not unconditionally independent, X is not a descendant of Y, Y is not a descendant of X and the two variables do not share a common ancestor. Such a pair of variable sets violates Common Cause. If there is a path between any $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, then that path must contain a collider. Thus, X and Y are d-separated by the empty set. By Lemma 1, this entails that if $\mathcal{N}$ satisfies CMC, then X and Y are independent conditional on the empty set, i.e. they are unconditionally independent. Since X and Y are, by stipulation, not unconditionally independent, $\mathcal{N}$ does not satisfy CMC. This shows that CMC implies Common Cause.

Finally, we can show that Screening Off and Common Cause imply CMC. The proof requires three of the "graphoid axioms", which Pearl and Paz (1985) prove follow from the probability axioms (in stating these axioms, "the union" of two variables is a shorthand for the union of singleton sets containing those variables):

- Decomposition: If X is unconditionally independent of the union of A and B, then X is unconditionally independent of A and X is unconditionally independent of B.

- Weak Union: If X is unconditionally independent of the union of A and B, then X is independent of A, given B and X is independent of B, given A.

- Contraction: If X is independent of A, given B and X is unconditionally independent of B, then X is unconditionally independent of the union of A and B.

Versions of these axioms also hold for smaller "universes" in which all independence relations have the same conditioning event, e.g. if X is independent of the union of

A and B conditional on C, then X is independent of the union of A conditional on C and X is independent of B conditional on C.

The proof proceeds as follows:

1. Y is not a descendant of X. (Assumption)

2. There either is or is not a directed path from Y to X. (Tautology)

3. If there is a directed path from Y to X, then either Y is a parent of X or Y is a non-parent ancestor of X. (Def. of 'Directed Path' and 'Parent')

4. If Y is a parent of X, then Y is trivially independent of X given $\mathcal{PA}_X$. (Def. of Conditional Independence)

5. If Y is a non-parent ancestor of X, then Y is independent of X conditional on $\mathcal{PA}_X$. (Screening Off)

6. If there is a directed path from Y to X, then Y is independent of X conditional on $\mathcal{PA}_X$. (Premises 3-5)

7. If there is no directed path from Y to X, then either X and Y are unconditionally independent or they are not. (Tautology)

8. If there is no directed path from Y to X and X and Y are unconditionally independent, then Y is either unconditionally independent of the union of X and $\mathcal{PA}_X$, or it is not. (Tautology)

9. If Y is unconditionally independent of the union of X and $\mathcal{PA}_X$, then Y is independent of X given $\mathcal{PA}_X$. (Weak Union).

10. If there is no directed path from Y to X, then there is no directed path between Y and any variables in the union of X and $\mathcal{PA}_X \setminus \mathcal{A}_Y$, where $\mathcal{A}_Y$ is the set of all ancestors of Y (Def. of Ancestor).

11. All common ancestors of Y and $\mathcal{PA}_X \setminus \mathcal{A}_Y$ are also common ancestors of X and Y. (Def. of Parent and Ancestor)

12. If there is no directed path between Y and any variables in the union of X and $\mathcal{PA}_X \setminus \mathcal{A}_Y$ and Y is not unconditionally independent of the union of X and $\mathcal{PA}_X$, then Y is independent of the union of X and $\mathcal{PA}_X \setminus \mathcal{A}_Y$ given the set $\mathcal{A}_{X,Y}$ of common ancestors of X and Y. (Common Cause, Premises 1 and 11)

13. If Y is independent of the union of X and $\mathcal{PA}_X \setminus \mathcal{A}_Y$ given $\mathcal{A}_{X,Y}$, then Y is independent of the union of X and $\mathcal{PA}_X \setminus \mathcal{A}_Y$ given the union of $\mathcal{A}_{X,Y}$ and $\mathcal{PA}_X \cap \mathcal{A}_Y$. (Follows from the fact that $\mathcal{PA}_X \cap \mathcal{A}_Y \subseteq \mathcal{A}_{X,Y}$)

14. If Y is independent of the union of X and $\mathcal{PA}_X \setminus \mathcal{A}_Y$ given the union of $\mathcal{A}_{X,Y}$ and $\mathcal{PA}_X \cap \mathcal{A}_Y$, then Y is independent of X given the union of $\mathcal{A}_{X,Y}$, $\mathcal{PA}_X \setminus \mathcal{A}_Y$ and $\mathcal{PA}_X \cap \mathcal{A}_Y$. (Weak Union)

15. If Y is independent of X given the union of $\mathcal{A}_{X,Y}$, $\mathcal{PA}_X \setminus \mathcal{A}_Y$ and $\mathcal{PA}_X \cap \mathcal{A}_Y$, then Y is independent of X, given the union of $\mathcal{PA}_X$ and $\mathcal{A}_{X,Y}$. (Def. of Set Difference, Union and Intersection)

16. If Y is independent of X, given the union of $\mathcal{PA}_X$ and $\mathcal{A}_{X,Y}$, and Y is unconditionally independent of X, then Y is unconditionally independent of X, $\mathcal{PA}_X$ and $\mathcal{A}_{X,Y}$. (Contraction)

17. If Y is unconditionally independent of the union of X, $\mathcal{PA}_X$ and $\mathcal{A}_{X,Y}$, then Y is unconditionally independent of the union of X and $\mathcal{PA}_X$. (Decomposition)

18. If Y is unconditionally independent of the union of X and $\mathcal{PA}_X$, then Y is independent of X, given $\mathcal{PA}_X$. (Weak Union)

19. If there is no directed path from Y to X and Y and X are unconditionally independent, then then Y is independent of X, given $\mathcal{PA}_X$. (Premises 8-16)

20. If there is no directed path from Y to X and Y and X are not unconditionally independent, then the union of X and $\mathcal{PA}_X$ is independent of Y, given $\mathcal{A}_{X,Y}$. (Common Cause, Premise 1)

21. If the union of X and $\mathcal{PA}_X$ is independent of Y, given a common ancestor $\mathcal{A}_{X,Y}$, then Y is independent of X given the union of $\mathcal{PA}_X$ and $\mathcal{A}_{X,Y}$. (Weak Union)

22. X is independent of $\mathcal{A}_{X,Y}$, given $\mathcal{PA}_X$. (Screening Off)

23. If Y is independent of X given the union of $\mathcal{PA}_X$ and $\mathcal{A}_{X,Y}$ and X is independent of $\mathcal{A}_{X,Y}$, given $\mathcal{PA}_X$, then X is independent of the union of Y and $\mathcal{A}_{X,Y}$, given $\mathcal{PA}_X$. (Contraction)

24. If X is independent of the union of Y and $\mathcal{A}_{X,Y}$, given $\mathcal{PA}_X$, then Y is independent of X given $\mathcal{PA}_X$. (Decomposition)

25. If there is no directed path from Y to X and Y and X are not unconditionally independent, then Y is independent of X given $\mathcal{PA}_X$. (Premises 18-22)

26. If there is no directed path from Y to X then Y is independent of X given $\mathcal{PA}_X$. (Premises 18, 23)

27. If Y is not a descendant of X, then Y is independent of X given $\mathcal{PA}_X$. (Premises 1-24)

This shows that Screening Off and Common Cause imply CMC. □

1.7.3 Proof of Proposition 2

Proof. Assume $p(y_j|do(x_i)) > p(y_j)$. Let φ be the set of values of Y other than y_j. This implies that $1 - p(y_j|do(x_i)) < p(\varphi)$, which implies in turn that there is a value y_l such that $p(y_l|do(x_i)) < p(y_l)$. Suppose that $\mathcal{PA}_X$ has the set of possible n-tuples of values $\{\mathbf{pa_{X1}}, \mathbf{pa_{X2}}, ..., \mathbf{pa_{Xq}}\}$. It is well known (see Pearl et al. 2016, p. 59) that we can derive the following:

$$p(y_l|do(x_i)) = \sum_{t=1}^{q} p(y_l|x_i, \mathbf{pa_{Xt}})p(\mathbf{pa_{Xt}}) < p(y_l) \tag{1.42}$$

The law of total probability implies the following, where X has n values:

$$p(y_l) = \sum_{k=1}^{n}\sum_{t=1}^{q} p(y_l|x_k, \mathbf{pa_{Xt}})p(x_k, \mathbf{pa_{Xt}}) \tag{1.43}$$

This implies that there exists a value x_k such that:

$$\sum_{t=1}^{q} p(y_l|x_i, \mathbf{pa_{Xt}})p(\mathbf{pa_{Xt}}) < p(y_l) < \sum_{t=1}^{q} p(y_l|x_k, \mathbf{pa_{Xt}})p(x_k, \mathbf{pa_{Xt}}) \tag{1.44}$$

The fact that $p(x_i) > 0$ implies that $p(x_k) < 1$, which implies in turn that $p(x_k, \mathbf{pa_{Xt}}) < p(\mathbf{pa_{Xt}})$ for all $\mathbf{pa_{Xt}}$ with positive probability. It follows from this that:

$$\begin{aligned} p(y_l) &< \sum_{t=1}^{q} p(y_l|x_k, \mathbf{pa_{Xt}})p(x_k, \mathbf{pa_{Xt}}) \\ &< \sum_{t=1}^{q} p(y_l|x_k, \mathbf{pa_{Xt}})p(\mathbf{pa_{Xt}}) = p(y_l|do(x_k)) \end{aligned} \tag{1.45}$$

Which immediately implies that $p(y_l|do(x_k)) > p(y_l)$. □

1.7.4 Proof of Proposition 3

Proof. Consider a Bayes net $\mathcal{N}' = \langle \mathcal{V}', \mathcal{E}', p(\cdot) \rangle$, with $C' \in \mathcal{V}'$, where $C' = c'_l$ explains why $E' = e'_s$. This explanatory fact implies that $p(e'_s|do(c'_l)) > p(e'_s)$. Define an equivalence relation $\sim$ over the value set of C' such that $c'_l \sim c'_u$ if and only if either:

a) $p(e'_s|do(c'_l)) > p(e'_s)$ and $p(e'_s|do(c'_u)) > p(e'_s)$, or

b) $p(e'_s|do(c'_l)) \leq p(e'_s)$ and $p(e'_s|do(c'_u)) \leq p(e'_s)$.

Let the value set of C be a quotient set of the value set of C' according to this equivalence relation. Define a Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ such that $\mathcal{V} = \{C\} \cup \mathcal{V}' \setminus \{C'\}$ and such that, for any edge $\langle V', C' \rangle \in \mathcal{E}'$ or $\langle C', V' \rangle \in \mathcal{E}'$, there is an edge $\langle V', C \rangle \in \mathcal{E}$ or $\langle C, V' \rangle \in \mathcal{E}$; all other elements of $\mathcal{E}'$ are included in $\mathcal{E}$. $\mathcal{N}$ trivially satisfies (1) and (4); we now show that it satisfies (2), (3), and (5) as well.

Beginning with (2), let $\mathcal{X} = \mathcal{V}' \setminus \{C'\} = \mathcal{V} \setminus \{C\}$ and let $\mathcal{X} = \mathbf{x}_i$ denote some assignment of values to all the variables in $\mathcal{X}$. If C' has q values and C has m values, then $p(\mathbf{x}_i) = \sum_{l=1}^{q} p(\mathbf{x}_i|do(c'_l))$ and $p(\mathbf{x}_i) = \sum_{j=1}^{m} p(\mathbf{x}_i|do(c_j))$. For each value c_j of C, there is a set of values $\{c'_1, c'_2, ..., c'_\alpha\}$ that are all $\sim$-related to each other and are mapped to c_j in the coarsening from C' to C, such that $\sum_{l=1}^{\alpha} p(\mathbf{x}_i|do(c'_l)) \frac{p^*(do(c'_l))}{p^*(do(c_j))} = p(\mathbf{x}_i|do(c_j))$, where probabilities are assigned to interventions in keeping with the procedure defined in Section 1.2.3, and otherwise arbitrarily. Since the value set of C is a partition of the value set of C', this equation generalizes so that $\sum_{l=1}^{n} p(\mathbf{x}_i|do(c'_l)) = \sum_{j=1}^{m} p(\mathbf{x}_i|do(c_j))$, and therefore $p(\mathbf{x}_i) = p(\mathbf{x}_i)$. This shows that condition (2) is satisfied.

We now move to condition (3). To show that $\mathcal{N}$ satisfies CMC, let $X' \in \mathcal{V}'$ and $Y' \in \mathcal{V}'$ be two variables such that Y' is not a descendant of X' in $\mathcal{N}'$. If $X' \neq C'$, $Y' \neq C'$, and $C' \notin \mathcal{PA}_{X'}$, then the supposition that $\mathcal{N}'$ satisfies CMC, along with the truth of conditions (1) and (2), implies that X' and Y' are independent, given $\mathcal{PA}_{X'}$, in $\mathcal{N}$.

If $C' = X'$, then the fact that $\mathcal{N}'$ satisfies CMC implies that for any values c'_l, y'_o, and n-tuple of values $\mathbf{pa}_{C'}$ of the variables in $\mathcal{PA}_{C'}$, $p(c'_l|y'_o, \mathbf{pa}_{C'}) = p(c'_l|\mathbf{pa}_{C'})$. If C is a coarsening of C', then for each value c_j, each conditional probability $p(c_j|y'_o, \mathbf{pa}_C)$ and $p(c_j|\mathbf{pa}_C)$ is a sum of terms of the form $p(c'_l|y'_o, \mathbf{pa}_{C'})$ and $p(c'_l|\mathbf{pa}_{C'})$, respectively. Thus, if the latter pair of terms are equal for all triples $(c'_l, y'_o, \mathbf{pa}_{C'})$, then the former pair of terms are equal for all values $(c_j, y'_o, \mathbf{pa}_C)$. Thus, C is independent of its non-descendants, given its parents, in $\mathcal{N}$.

If $C' = Y'$, then the fact that N' satisfies CMC implies that for any values c'_l, x'_o, and set of values $\mathbf{pa}_{\mathbf{X}'}$ of the variables in $\mathcal{PA}_{X'}$, $p(x'_o|c'_l, \mathbf{pa}_{\mathbf{X}'}) = p(x'_o|\mathbf{pa}_{\mathbf{X}'})$. These conditional probabilities can be expressed as the following ratios:

$$p(x'_o|c'_l, \mathbf{pa}_{\mathbf{X}}) = \frac{p(x'_o, c'_l, \mathbf{pa}_{\mathbf{X}'})}{p(c'_l, \mathbf{pa}_{\mathbf{X}'})} \tag{1.46}$$

$$p(x'_o|\mathbf{pa}_{\mathbf{X}'}) = \frac{p(x'_o, \mathbf{pa}_{\mathbf{X}})}{p(\mathbf{pa}_{\mathbf{X}'})} \tag{1.47}$$

If C is a coarsening of C', then for each value c_j, each joint probability $p(x'_o, c_j, \mathbf{pa}_{\mathbf{X}'})$ and $p(c_j, \mathbf{pa}_{\mathbf{X}'})$ is equal to a sum of joint probabilities of the form $p(x'_o, c'_l, \mathbf{pa}_{\mathbf{X}'})$ and $p(c'_l, \mathbf{pa}_{\mathbf{X}'})$, respectively. Thus, if $p(x'_o|c'_l, \mathbf{pa}_{\mathbf{X}'}) = p(x'_o|\mathbf{pa}_{\mathbf{X}'})$ for all triples $(c'_l, x'_o, \mathbf{pa}_{\mathbf{X}'})$, then $p(x'_o|c_j, \mathbf{pa}_{\mathbf{X}'}) = p(x'_o|\mathbf{pa}_{\mathbf{X}'})$ for all triples $(c_j, x'_o, \mathbf{pa}_{\mathbf{X}'})$. Thus, if C is a non-descendant of X' in $\mathcal{N}$, then X' is independent of C, given X''s parents, in $\mathcal{N}$. The immediately preceding analysis could be repeated if $C \in \mathcal{PA}_{X'}$, to show that any variable X' is independent of its non-descendants in $\mathcal{N}$, given its parents, when those parents include C. Together, these results show that $\mathcal{N}$ satisfies CMC. Minimality can be achieved by stipulation, by simply removing any edges that are not necessary for $\mathcal{N}$ to satisfy CMC.

Finally, we can show that (5) is true. Suppose that $C' = c'_l$ implies that $C = c_j$. In other words, c'_l is mapped to c_j in the coarsening function from the value set of C' to the value set of C. We know that $p(e'_s|do(c'_l)) > p(e'_s)$, and c'_l is $\sim$-related to all and only those values of C' such that conditioning on an intervention bringing about those values increases the probability that $E' = e'_s$, relative to its marginal probability. Thus, the conditional probability $p(e'_s|do(c_j))$ is a sum of terms of the form $p(e'_s|do(c'_l))$, each of which is such that $p(e'_s|do(c'_l)) > p(e'_s)$. This implies that $p(e'_s|do(c_j)) > p(e'_s)$, and therefore that $C = c_j$ explains $E' = e'_s$. □

2

Explanatory Goodness and Explanatory Power

2.1 Introduction

In the previous chapter, I used the formal machinery of Bayes nets to argue that in many cases the same event can be causally explained at different levels of granularity. In this chapter, I begin to consider how we might go about selecting an *optimal* level of granularity for causal explanation, given that several levels are possible. Specifically, I examine a series of measures of *explanatory power* put forward in the literature. One might hope that these measures would be helpful in selecting an optimal level of granularity for causal explanation. However, in this chapter, I show that applying these measures in this way would lead to judgments that are at odds with scientific practice. Therefore, such an application of these measures ought to be avoided.

Schupbach and Sprenger (2011), Crupi and Tentori (2012), and Eva and Stern (forthcoming) all propose competing Bayesian measures of the power with which some hypothesis explains some evidence. Schupbach and Sprenger argue that their measure of explanatory power "clarif[ies] the conditions under which hypotheses are judged to provide strong versus weak explanations of some proposition", and also "clarif[ies] the meaning of comparative explanatory judgments such as 'hypothesis $[H_1]$ provides a better explanation of [some] fact than does hypothesis $[H_2]$'" (2011, p. 106). These claims suggest that Schupbach and Sprenger take their measure of explanatory power to track the overall goodness of an explanation, which is sometimes called its "explanatory depth" (see Hitchcock and Woodward 2003, Strevens 2008, and Weslake 2010). Since Crupi and Tentori attempt to critique Schupbach and Sprenger's original measure while taking a similar formal approach, I read them as broadly committed to the claim that their measure also tracks overall explanatory goodness or depth. As for Eva and Stern, I will note in what follows that they acknowledge the limitations of their measure as a measure of overall explanatory goodness, although they do not canvas the full set of cases in which their measure may fail in this way.

My aim in this chapter is to argue that each of these measures cannot be measures of the overall goodness or depth of an explanation. My argument takes as its starting point a series of formal results showing that all three measures of explanatory power essentially track the accuracy with which knowledge of a hypothesis allows us to predict the occurrence of some evidence, where accuracy is measured using the Brier score. I then show that there are many instances in science in which this predictive accuracy can be sacrificed in favor of other explanatory virtues, in order to achieve a better explanation of some phenomenon. I focus here on cases in which, by moving to a more coarse-grained description of some hypothesis, scientists are able to obtain a better explanation, despite sacrificing accuracy. Thus, I conclude that the explanatory power measures proposed in the literature do not measure overall explanatory goodness.

Before moving forward, it is worth addressing a salient concern with respect to the framing of my argument. One could argue that the term 'explanatory power' has the same extension as 'explanatory goodness', such that my argument shows that the proposed measures do not track what they claim to track. This is essentially a terminological dispute. For the sake of disambiguation, going forward I will grant that there is a feature of explanations called 'explanatory power' that the proposed measures may track, but argue that none of these measures necessarily track the overall goodness of an explanation.

The plan for this chapter is as follows. In Section 2.2, I provide an overview of the three measures of explanatory power considered here. In Section 2.3, I state the formal results showing that all three measures of explanatory power are ordinally equivalent to the accuracy with which the explanatory hypothesis allows us to predict the evidence being explained. In Section 2.4, I discuss the nature of explanatory goodness. In Section 2.5, I show that explanatory goodness is not always ordinally equivalent to the the accuracy with which an explanans predicts its explanandum, and therefore not always ordinally equivalent to any of the proposed measures of explanatory power. In Section 2.6, I clarify the relationship between the granularity of an explanatory hypothesis, explanatory power, explanatory goodness, and the accuracy with which an explanatory hypothesis predicts the evidence that it explains. In Section 2.7, I discuss two possible alternative measures to the three considered here—one due to Ströing (2018) and one that I introduce in the following chapter—that do not run afoul of the problems that I have raised. While I am agnostic here as to which measure is to be preferred, I note that both of these proposed functions have additional arguments beyond those that are common to Schupbach and Sprenger, Crupi and Tentori, and Eva and Stern's measures. In Section 2.8, I offer concluding remarks.

2.2 Measures of Explanatory Power

I begin with Schupbach and Sprenger's (2011) measure. Let Ω be a set of possible worlds, and let $\mathcal{A}_\Omega$ be an algebra on Ω, i.e. a set of subsets of Ω that is closed under complement, union, and intersection. Let $p(\cdot)$ be a probability distribution defined over $\mathcal{A}_\Omega$. The proposition $H \in \mathcal{A}_\Omega$ denotes the set of possible worlds in which some hypothesis is true. The proposition $E \in \mathcal{A}_\Omega$ denotes the set of possible worlds in which some evidence is observed. It is assumed that $p(E) \neq 0$. Schupbach and Sprenger claim that the power with which H explains E is measured by the following function:

$$\varepsilon_{ss}(H, E) = \frac{p(H|E) - p(H|\neg E)}{p(H|E) + p(H|\neg E)} \in [-1, 1] \tag{2.1}$$

A high value of $\varepsilon_{ss}(H, E)$ represents a high degree of explanatory power for H with respect to E, and a low value represents a low degree of explanatory power for H with respect to E. If $\varepsilon_{ss}(H, E) = 0$, then H has neutral explanatory power with respect to E. Schupbach and Sprenger argue for the appropriateness of this measure by defining a set of conditions that any measure of explanatory power ought to satisfy, and then proving that their measure in fact satisfies these conditions.

To understand Schupbach and Sprenger's measure, it is crucial to understand their intended interpretation of the probability function $p(\cdot)$ with respect to the propositions H and E. Note that the two probabilistic terms in their measure can be re-written as follows: $p(H|E) = \frac{p(E|H)p(H)}{p(E)}$, $p(H|\neg E) = \frac{(1-p(E|H))p(H)}{1-p(E)}$. They argue that $p(E|H)$ represents some agent's degree of belief that E is true, given the supposition that the hypothesis H is true. By contrast, $p(H)$ represents the degree of belief that an agent would have in the truth of H, if they did not know the truth value of any other propositions. Finally, $p(E)$ is the degree of belief that the agent would have in the truth of E, if they did not know the truth value of any other propositions. All of these interpretations contain some counterfactual content, since, in fact, the agent *does know* that E is true. After all, E describes some actual observed evidence, which H explains to some degree. The upshot here is that we should interpret probabilities in Schupbach and Sprenger's measure not as representing any actual agent's degree of belief, but rather as the degree of belief that an idealized agent would have under certain counterfactual conditions. Other attempts to measure explanatory power make similar assumptions. In what follows, I grant that these assumptions are justified.

Crupi and Tentori (2012) take issue with Schupbach and Sprenger's measure. They show that if E^* is some proposition that is probabilistically independent of E, and if $p(E|H) \leq p(E)$, then $\varepsilon_{ss}(H, E) < \varepsilon_{ss}(H, E \wedge E^*)$. Thus, if conditionalizing on

H lowers the probability of E, then H explains the conjunction of E and some irrelevant proposition E^* more powerfully than it explains E alone, according to Schupbach and Sprenger's measure. Crupi and Tentori take this to be an undesirable result, and therefore propose the following alternative measure of explanatory power, which is not subject to the same worry:

$$\varepsilon_{ct}(H,E) = \begin{cases} \frac{p(E|H)-p(E)}{1-p(E)} & \text{if } p(E|H) \geq p(E) \\ \frac{p(E|H)-p(E)}{p(E)} & \text{if } p(E|H) < p(E) \end{cases} \in [-1,1] \tag{2.2}$$

My aim here is not to adjudicate between these two measures, but rather to show that they both stand in the same ordinal equivalence relationship to a Brier score measure of accuracy. See Cohen (2016) for a defense of Schubach and Sprenger's measure against Crupi and Tentori's arguments.

Finally, Eva and Stern (forthcoming) propose a third measure of explanatory power. Their measure is not so much a critique of Schupbach and Sprenger's measure as it is an application of their measure to a specific formal context, namely the context of causal Bayesian networks. As such, their measure is most directly relevant to my arguments in the rest of this dissertation. Recall that a Bayes net is a triple $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$, where $\mathcal{V}$ is a set of random variables such that each is defined over the same set of possible worlds Ω, $\mathcal{E}$ is a set of ordered pairs or 'edges' defined over $\mathcal{V}$, and $p(\cdot)$ is a joint distribution defined over the Cartesian product of each measure space of each random variable in $\mathcal{V}$. If there is an edge from a variable X to Y, then we say that X is a parent of Y. If there is a chain of edges beginning at X and terminating at Y, then we say that Y is a descendant of X. A Bayes net satisfies the *Causal Markov Condition*, which says that any variable is probabilistically independent of its non-descendants, conditional on its parents. This condition allows us to calculate, for any variables X and Y and any values x_i and y_j, the probability $p(x_i|do(y_j))$, where $do(y_j)$ denotes that the variable Y is set to the value y_j via an *intervention*, or exogenous change to the system that is independent of any other variables in the network other than the variable being intervened upon and its descendants. Since the values of these probabilities depend on the graphical structure of the Bayes net in which the variables X and Y are embedded, calculating them requires that we know the relations $\mathcal{E}$ over the variables in a given Bayes net.

Against this formal background, Eva and Stern define the following measure of causal explanatory power. Let H be a random variable in some Bayes net whose values denote different possible hypotheses, and let E be a random variable whose values denote different possible observations. We can use the Bayes net in which

these variables are embedded to define the following measure of the power with which H taking some value h_j causally explains E taking some value e_i.

$$\varepsilon_{es}(h_j, e_i, \mathcal{N}) = \frac{p(do(h_j)|e_i) - p(do(h_j)|\neg e_i))}{p(do(h_j)|e_i) + p(do(h_j)|\neg e_i)} \in [-1, 1] \tag{2.3}$$

This measure has the same form as Schupbach and Sprenger's, but is explicitly embedded within the Bayes nets formalism. In particular, since the conditional probabilities of interventions $p(do(h_j)|e_i)$ and $p(do(h_j)|\neg e_i))$ can only be calculated once we know the structure of the Bayes net in which the variables H and E are embedded, the Bayes net $\mathcal{N}$ must be an argument of Eva and Stern's function. Let us assume that $\mathcal{N}$ is an "extended" Bayes net that includes intervention variables, as defined in Section 1.2.3. Thus, Eva and Stern's measure is defined in relation to a particular sort of statistical model of the relation between hypothesis and evidence. To a degree, this speaks in favor of their result. Unlike Schupbach and Sprenger's measure, which requires us to interpret probabilities as the degrees of belief of counterfactual agents, Eva and Stern's measure just imports its probabilities from some statistical causal model of the scenario in question. However, like both Schupbach and Sprenger and Crupi and Tentori's measure, Eva and Stern's measure is ordinally equivalent to the Brier score measure of the accuracy with which the hypothesis predicts the evidence. I turn to this result in the next section.

2.3 Power and Accuracy

The accuracy with which we can predict the occurrence of some evidence, given the supposition that a hypothesis is true, can be made formally precise via scoring functions. The most popular of these functions is the Brier score. The general form of the Brier score is as follows. Let $\mathcal{F} = \{F_1, F_2, ..., F_n\}$ be a partition of an outcome space Ω such that each element of $\mathcal{F}$ is in an algebra $\mathcal{A}_\Omega$. Let $O = \{o_1, o_2, ..., o_n\}$ be a set of numbers such that if the event represented by F_i actually occurs, then $o_i = 1$; otherwise, $o_i = 0$. The Brier score for a probability distribution $p(\cdot)$ over $\mathcal{A}_\Omega$ is given by the following equation:

$$\mathfrak{B}(p(\cdot), \mathcal{F}) = \frac{1}{n} \sum_{i=1}^{n} \left(p(F_i) - o_i\right)^2 \tag{2.4}$$

The Brier score takes a value in the interval $[0, 1]$, with a lower Brier score indicating a more accurate probability distribution over possible outcomes. For a comprehensive defence of the Brier score as an optimal measure of accuracy, see Joyce (2009) and Pettigrew (2016).

To apply this measure to the particular case of predicting the truth of some evidential proposition E, let $\mathcal{E} = \{E, I_1, I_2, \ldots, I_n\}$ be a partition of Ω such that each element of $\mathcal{E}$ is in the algebra $\mathcal{A}_\Omega$. If H is an element of $\mathcal{A}_\Omega$, then we can define a conditional probability $p(\cdot|H)$ to each element of $\mathcal{E}$.[1] Assuming that the proposition E is true, we can measure the accuracy of $p(\cdot|H)$ with respect to $\mathcal{E}$ using the following Brier score:

$$\mathfrak{B}(p(\cdot|H), \mathcal{E}) = \frac{1}{n+1}\left(\left(p(E|H) - 1\right)^2 + \sum_{i=1}^{n}\left(p(I_i|H) - 0\right)^2\right) \tag{2.5}$$

If, for any two hypotheses H^1 and H^2 in the algebra $\mathcal{A}_\Omega$, it is the case that $\sum_{i=1}^{n} p(I_i|H_1) - p(I_i|H_2) = p(E|H_1) - p(E|H_2)$, then the closer that the probability $p(E|H)$ is to one, the lower the Brier score will be, and therefore the more accurate the distribution $p(\cdot|H)$ will be with respect to the partition $\mathcal{E}$.[2] This fits with a commonsense notion of accuracy. Since we have assumed that E is true, the conditional probability distribution is deemed to be accurate with respect to the partition $\mathcal{E}$ to the extent that it assigns a high probability to E and a low probability to false elements of $\mathcal{E}$.

We can now prove two results showing a close connection between both Schupbach and Sprenger and Crupi and Tentori's measures of explanatory power. Let E, H_1, and H_2 be elements in some algebra $\mathcal{A}_\Omega$, and let $\mathcal{E}$ be a partition of the outcome space Ω such that $E \in \mathcal{E}$. The propositions $H_1 \in \mathcal{A}_\Omega$ and $H_2 \in \mathcal{A}_\Omega$ are competing hypotheses that explain some evidential hypothesis E. The following propositions are true:[3]

Proposition 4. *If E is true, and if for any two hypotheses H^1 and H^2 in the algebra $\mathcal{A}_\Omega$, $\sum_{i=1}^{n} p(I_i|H_1) - p(I_i|H_2) = p(E|H_1) - p(E|H_2)$, then $\varepsilon_{ss}(H_1, E) > \varepsilon_{ss}(H_2, E)$ if and only if $\mathfrak{B}(p(\cdot|H_1), \mathcal{E}) < \mathfrak{B}(p(\cdot|H_2), \mathcal{E})$.*

Proposition 5. *If E is true, and if for any two hypotheses H^1 and H^2 in the algebra $\mathcal{A}_\Omega$, $\sum_{i=1}^{n} p(I_i|H_1) - p(I_i|H_2) = p(E|H_1) - p(E|H_2)$, then $\varepsilon_{ct}(H_1, E) > \varepsilon_{ct}(H_2, E)$ if and only if $\mathfrak{B}(p(\cdot|H_1), \mathcal{E}) < \mathfrak{B}(p(\cdot|H_2), \mathcal{E})$.*

In other words, both Schupbach and Sprenger's measure and Crupi and Tentori's measure are ordinally equivalent to the Brier score measure of the accuracy with

[1] Although it is not the most elegant use of formalism, it should be clear from context where we are discussing the partition $\mathcal{E}$ and the set of edges in a Bayes net $\mathcal{E}$.

[2] The requirement that $\sum_{i=1}^{n} p(I_i|H_1) - p(I_i|H_2) = p(E|H_1) - p(E|H_2)$ is needed in order to ensure that $\frac{1}{n+1}\left(p(I_i|H) - 0\right)^2$ is always inversely proportional to $P(E|H)$.

[3] Note that I assume in the proofs of these propositions that E, H_1, and H_2 are all elements of a single algebra in a single probability space. Schupbach and Sprenger make this assumption in proofs of various propositions related to the properties of their measure (2011, p. 118). This assumption rules out the possibility that H_1 and H_2 are probabilistic hypotheses that posit different probability distributions.

which conditionalizing on a hypothesis allows us to predict the occurrence of some evidence. The more accurate this prediction, the more powerful the explanation, and vice-versa.

In order to discuss the relation between Eva and Stern's distinctly causal measure of explanatory power and predictive accuracy, we need to first import the Brier score accuracy measure into the Bayes nets formal context. To do this, let E and H be variables in a Bayes net $\mathcal{N}$ that take the values e_i and h_j respectively, and let E take n possible values. The causal accuracy with which conditionalizing on an intervention $do(h_j)$ allows us to predict the value of E is given by the following equation:

$$\mathfrak{B}_c(p(\cdot|do(h_j)), E, \mathcal{N}) = \frac{1}{n}(p(e_i|do(h_j)) - 1)^2 + \frac{1}{n}\sum_{k=1|i\neq k}^{n}(p(e_k|do(h_j)) - 0)^2) \tag{2.6}$$

This function measures how close the probability distribution over E comes to successfully predicting the actual value of E, once we conditionalize on an intervention such that $H = h_j$.

Given the similarity of Eva and Stern's measure of explanatory power and Schupbach and Sprenger's, it should not be surprising that there is a close connection between Eva and Stern's measure and the Brier score measure of accuracy given in equation (2.6). For a given Bayes net $\mathcal{N}$ with variables H^1, H^2, and E, the following proposition is true:

Proposition 6. *If $E = e_i$, and for any two variable values h_j^1 and h_l^2, it is the case that $\sum_{k=1|i\neq k}^{n} p(e_k|do(h_j^1)) - p(e_k|do(h_l^2)) = p(e_i|do(h_j^1)) - p(e_i|do(h_l^2))$, then it is the case that $\varepsilon_{es}(h_j^1, e_i, \mathcal{N}) > \varepsilon_{es}(h_l^2, e_i, \mathcal{N})$ if and only if $\mathfrak{B}_c(p(\cdot|do(h_j^1)), E, \mathcal{N}) < \mathfrak{B}_c(p(\cdot|do(h_l^2)), E, \mathcal{N})$.*

This proposition allows for the possibility $H^1 = H^2$. Thus, according to Eva and Stern's measure, comparisons of relative causal explanatory power track comparisons of relative causal predictive accuracy whether we are comparing: 1) the causal explanatory power of two variables in the same Bayes net taking different values, or 2) the causal explanatory power of the same variable in the same Bayes net taking different values. All that must be held fixed is the explanandum $E = e_i$.

These results show that the three measures of explanatory power necessarily track the accuracy with which the supposition that a given hypothesis is true allows us to predict the occurrence of some evidence. In the next two sections, I argue that this fact about these explanatory power measures implies that they cannot be

measures of the overall goodness or depth of an explanation. This is because the comparative goodness with which two different hypotheses explain some evidence is not necessarily ordinally equivalent to the accuracy with which each hypothesis allows us to predict the occurrence of the evidence.

2.4 The Nature of Explanatory Goodness

I take it for granted here that, for any given body of evidence, there is not a unique hypothesis such that only that hypothesis explains the evidence in question. Rather, multiple hypotheses can explain the same evidence. However, this does not entail that all explanations provide equally good explanations of a given body of evidence. The following quote from Hitchcock and Woodward (2003) exemplifies this view:

> Some explanations are deep and powerful: Newton's explanation of the tides, Maxwell's explanation of the propagation of light, Einstein's explanation of the advance of the perihelion of Mercury. Other explanations, while deserving of the name, are superficial and shallow: Bob lashed out at Tom because he was angry, the car accelerated because Mary depressed the gas pedal with her foot, the salt dissolved because it was placed in water. We take this intuition to be very natural and widely shared (p. 181).

This gives us some intuitive examples of deeper and shallower explanations, without giving much of an account of explanatory depth itself. The common thread running through these examples, and others like them, is that explanatory depth captures something about the overall superiority of an explanation as compared to other explanations of the same phenomenon. Weslake (2010) expresses a similar sentiment, writing: "by explanatory depth, I mean a measure in terms of which explanations can be assessed according to their explanatory value" (p. 273). On this basis, I stipulate that 'explanatory depth' has the same extension as 'explanatory goodness'.

Hitchcock and Woodward (2003) suggest that explanatory goodness is closely related to the degree to which an explanation allows us to understand why some phenomenon occurred. They discuss a case in which the conjunction of Galileo's laws of motion and the Boyle-Charles law describing the thermodynamic behavior of gases provides an explanation of the time it takes an object to fall from some height. They claim that this explanation is "no deeper" than an explanation that cites only the Galilean laws, because "the conjunctive law does not increase our understanding" of why the object falls for the length of time that it does (p. 190). Strevens (2008) advances a similar conception of the relation between explanation

and understanding, claiming that a deep explanation "confers full understanding of the explanandum" (p. 154). For the sake of argument, I take on board the point that our intuitions about which explanations improve our understanding of a phenomenon track, to some degree, the comparative goodness or depth of those explanations. However, I do not mean to suggest that explanatory goodness simply *is* the degree to which an explanation provides understanding. Rather, it may be that the sense of understanding that certain explanations inspire is the result of a number of features of an explanation.

While Hitchcock and Woodward's examples paint an evocative picture of the notion of explanatory goodness, they obviously lack formal precision. Thus, Schupbach and Sprenger, Crupi and Tentori, and Eva and Stern are all justified in their attempt to measure explanatory goodness in a mathematically precise way. However, I will argue that their measures capture, at most, a small portion of what we talk about when we talk about explanatory depth. This is because explanatory goodness does not necessarily track the accuracy with which a hypothesis predicts some evidence, whereas these explanatory power measures do.

2.5 Explanatory Goodness, Accuracy, and Power

Consider a scenario involving the diffusion of a droplet of dye in a container of water. We can describe this scenario in at least three different ways, where each of these three descriptions furnishes a different explanation of why a given particle of the dye is within a 9mm Euclidean distance from its origin.[4] Some of these explanations are better than others, such that we are able to generate a rank-ordering of cases according to their overall explanatory goodness. In each of these descriptions, we can interpret conditional probabilities either as ordinary conditional probabilities or as causal conditional probabilities, i.e. conditional probabilities in which the conditioning event is some intervention. For the sake of argument, I assume going forward that these two probabilities are equal to each other.

> **Description A**: Suppose that we drop some blue dye into a container of water at $t = 0\text{s}$. If we assume that the diffusion of the dye in water can be modelled by a multivariate standard normal distribution, that the diffusion of the dye is spherical, and that the diffusion constant for the dye in water (largely a function of the temperature of the water) is $1\frac{\text{mm}^2}{\text{s}}$, then the probability that an given dye particle is within a 9mm Euclidean distance from its origin at $t = 1\text{s}$ is approximately .97.

[4] A similar sort of example is alluded to in List and Pivato (2015b).

Explanation A: The dye particle is within 9mm of the origin at $t = 1\text{s}$ because the diffusion constant is $1\frac{\text{mm}^2}{\text{s}}$.

Description B: Using a microphysical, deterministic model of the water container that represents not just the temperature of the water and certain physical properties of the dye, but also the directional velocity of each water molecule at the time that the droplet enters the container, we can assign probability 1 or 0 to the event that a given dye molecule is within a 9mm Euclidean distance from its origin at $t = 1\text{s}$.

Explanation B: The dye particle is within 9mm of the origin at $t = 1\text{s}$ because the microphysical state of the system is x.

Description C: Using a highly coarse-grained model of the container that specifies whether the water is hot or cold, we determine that since the water is cold, the probability that the particle is within a 9mm Euclidean distance from its origin at $t = 1\text{s}$ is approximately $.65$.

Explanation C: The dye particle is within 9mm of the origin at $t = 1\text{s}$ because the water is cold.

There is a case to be made that Explanation A provides the best explanation of why the dye is within 9mm of the origin. As List and Pivato (2015b) note, the $.97$ probability assigned to the event that a dye droplet is within 9mm of its origin is derived via a special application of the *heat equation*, a differential equation that has great descriptive power for all sorts of processes in which a substance spreads out over time. As Weslake (2010) argues, deep explanations subsume many token instances of causal dependence under the same general patterns; Explanation A satisfies this desideratum.[5]

By contrast, Explanation B has the negative characteristic of being what Weslake (2010), following Garfinkel (1981), calls *hyperconcrete*. A hyperconcrete explanation is an explanation in which the explanandum event is described in so fine-grained a way that it is unlikely that the event, so described, will ever occur more than once. Given the number of degrees of freedom for the possible microstates of a macroscopic container of water, it is unlikely that two such containers will ever be in the same micro-state twice. If explanatory depth increases when an explanatory model allows us to subsume the different token instances of a given explanation under the same explanatory pattern, then it stands to reason that hyperconcrete explana-

[5] Arguably, the idea that deeper explanations subsume more events under a general pattern goes back as far as Hempel and Oppenheim's (1948) covering-law model of explanation.

tions of a given phenomenon are decidedly less deep than their non-hyperconcrete counterparts. Thus, we can rationalize the intuition that Explanation A is a deeper explanation than Explanation B, with respect to why the dye particle was less than 9mm from the origin.

Not everyone will be convinced by the claim that Explanation A is better than Explanation B. After all, the development of a deterministic, micro-level model of the diffusion of dye in water would be a scientific achievement, offering insight into the fine-grained structure of matter. While I will hold for the sake of argument that Explanation A is deeper than Explanation B, cases from the social sciences may show more clearly the sense in which hyperconcrete explanations can be undesirable. To illustrate, suppose that we are trying to explain the outcome of a recent election. In one case, we explain the outcome by citing the percentage of each key demographic in a population that voted for each candidate. In a second case, we explain the election outcome by citing how every voter in every constituency actually voted. In both cases, we have an explanation of the second election result. However, the second explanation is hyperconcrete and minutely detailed. It seems to miss the forest for the trees, and convey little understanding of why a given result occurred.

At first glance, empirical work by Bechlivanidis et al. (2017) would seem to cast doubt on my case for the explanatory superiority of Explanation A over Explanation B. They find that in a host of experimental scenarios, subjects preferred fine-grained explanations to coarse-grained explanations of the same event, even when the two explanatory hypotheses were equally good guides to predicting the explanandum. For instance, subjects are told that a student failed a test after studying thirty minutes per day in the time leading up to the test. They are also told that test-takers usually need five-hours of study each day to pass the test. In this case, subjects felt that the explanation 'the student failed because they studied thirty minutes per day' was a better explanation than 'the student failed because they studied less than five hours per day'. These results seem to indicate that the goodness of an explanation is closely tied to the level of detail with which the explanans is described. If true, this would seem to rule in favor of Explanation B over Explanation A.

In response, I argue that there is a crucial disanalogy between the kinds of explanatorily beneficial coarsenings that are found in scientific explanations and the coarsenings used in Bechlivandis et al.'s examples. In Explanation B, the explanans is decidedly more *complex* than the explanans in Explanation A. A formal description of the directional velocity of every water molecule will be significantly longer than a description of the dye's diffusion constant in water. The same is true of a description of each voter's behavior in an electorate, as compared to a description of broader voting trends. The same cannot be said of the explanatory phrases 'studying for thirty minutes' and 'studying less than five hours'; the two descriptions have roughly

equal complexity.[6] Thus, while Bechlivandis et al.'s findings do seem to debunk the thesis that explanatory depth is solely a function of the coarseness with which an explanans is described, they do not rule out the idea that a lack of complexity is a good-making feature of an explanation.

At the same time, I take it that Explanation A is better than Explanation C. Even if we stipulate that, within a given model, the distinction between hot and cold water can be precisely defined along well-established boundaries, the claim that the water is cold does not have a precise meaning within any successful scientific theory. Further, in contrast with an explanation that cites the diffusion constant of the dye in the water, this explanation tells us only of the effect of heating the water so that it goes from hot to cold. It tells us nothing about what might happen under a large range of other feasible interventions on the temperature of the water. Hitchcock and Woodward (2003) take the ability of an explanation to tell us what would happen under a wide range of possible interventions to be a good-making feature of an explanation, and it is just this feature that Explanation C lacks. This contrast between Explanation A and Explanation C shows that it is possible for an explanation to be shallow not because it is hyperconcrete, but rather, because it is too coarse-grained to describe what would change about the explanandum under a sufficiently wide range of possible interventions.

The upshot of this discussion is that *the explanatory goodness of a given explanation is not necessarily monotonically associated with the accuracy with which conditioning on an explanatory hypothesis allows us to predict the occurrence of some evidence*. In Explanation A, if a dye droplet is within 9mm of the origin, then the accuracy with which this can be predicted, given the initial conditions described in the setup (or given an intervention setting them to their actual values), is $\frac{1}{2}((.97-1)^2+(.03-0)^2) = .009$. In Explanation B, if a dye droplet is within 9mm of the origin, then the accuracy with which this can be predicted, given the initial microphysical conditions specified by the model (or given an intervention setting them to their actual values), is 0, since the model is deterministic. In Explanation C, if a dye droplet is within 9mm of the origin, then the accuracy with which this can be predicted, given that the water in the container is cold (perhaps due to an intervention), is $\frac{1}{2}((.65-1)^2+(.35-0)^2) \approx .12$. Thus, Explanation A has a middling level of accuracy, even though it provides the deepest explanation of the diffusion of the dye in water. In all of these cases, I have assumed that, if the accuracy measure in question is the causal accuracy measure $\mathfrak{B}_c(\cdot)$, then the intervened-upon variable has only two possible values, one of which is the relevant explanans value.

[6] Although the formal details are not important to my discussion here, I have in mind Kolmogorov's (1963) algorithmic measure of complexity.

Note the importance of the qualifier 'necessarily' in the claim that explanatory goodness is not necessarily monotonically related to the accuracy with which conditioning on an explanatory hypothesis allows us to predict its effect. It may be that within some class of explanations, the best explanation of some evidence is also the one such that the explanatory hypothesis allows us to most accurately predict the evidence in question. However, the examples above demonstrate that the best explanation of some phenomenon need not be derived from an explanatory model that maximizes the accuracy with which the explanans allows us to predict the explanandum. Thus, while explanatory goodness and explanatory power may be ordinally equivalent in some classes of explanations, the goodness and power of an explanation can come apart in other sets of explanations.

By contrast, the results in the previous section demonstrate that explanatory power—as measured by Schupbach and Sprenger, Crupi and Tentori, and Eva and Stern—is necessarily ordinally equivalent to the accuracy with which conditionalizing on an explanatory hypothesis allows us to predict some evidence. So all three measures will rank Explanation B as the most powerful, followed by Explanation A, and finally Explanation C. I take this result to illustrate a fundamental problem for these measures. Schupbach and Sprenger argue that their measure of explanatory power offers normative guidance with respect to how we *ought* to rank explanations with respect to their overall explanatory goodness, and I read other proposals for a measurement of explanatory power as having similar aims (2011, pp. 117-8). The examples of explanations given above show that this normative recommendation is out of keeping with scientific practice in at least some respects, as there are contexts in which generality and a lack of complexity are deemed to be explanatory virtues. I take these results as evidence that these explanatory power measures, in virtue of necessarily tracking the accuracy with which an explanatory hypothesis predicts some evidence, do not always measure the overall goodness or depth of an explanation.

As further evidence of the rift between explanatory goodness and predictive accuracy, Yarkoni and Westfall (2017) argue that in psychology, scientists often need to trade off explanatory success for predictive success, and vice-versa. They argue that "researchers must make a conscious choice: to explain or to predict" (p. 2). Yarkoni and Westfall motivate this distinction between explanation and prediction by citing statistical machine learning models that offer better predictions with respect to the psychological traits of test subjects than do more explanatory models. On this basis, they argue that contemporary psychology, which they claim largely prioritizes explanation over prediction, should adopt a more prediction-centric approach. While this argument is tangential to my claims here, Yarkoni and Westfall's argument does illustrate that practicing scientists sometimes embrace the claim that a more explanatory hypothesis can sometimes be less predictively accurate with respect to

its explanandum than some other hypothesis. This serves as evidence for the claim that if an explanatory power measure necessarily tracks accuracy, then it cannot be a measure of overall explanatory goodness.

Before moving forward, it is worth noting that Eva and Stern do anticipate some of the objections to the explanatory measures discussed here, including their own measure. They consider the case of someone trying to explain why a local football team lost a game (p. 24). The explanation 'they were losing by fifty points at the start of the fourth quarter', seems shallow and superficial, while the explanation 'their best player was injured at the start of the game', seems deeper and better. However, under some reasonable assumptions, Eva and Stern's measure will render the verdict that the first explanation of why the team lost is the better one. They conclude that their measure does not necessarily track the depth or goodness of an explanation, while noting that the same can be said of Schupbach and Sprenger's measure. However, this conclusion is limited in its scope. First, Eva and Stern do not diagnose what I take to be the fundamental limitation of these measures, which is that they are essentially measures of predictive accuracy. Second, they note that their measure fails to track explanatory goodness in cases where we feel that the better explanation is one that cites a more distant causal ancestor of the explanandum. This sort of case is different from the case that I have considered here, where the better explanation of some evidence does not cite a more distant causal ancestor, but instead describes the same causal factor in a more parsimonious, less complex way.

2.6 Clarifications Regarding Granularity and Accuracy

In the cases above, I demonstrated how explanatory goodness and predictive accuracy can come apart by showing how a more coarse-grained description of an explanatory hypothesis can sometimes result in a better explanation of some evidence, even at the cost of the accuracy with which the hypothesis predicts the evidence. However, I do not mean for this argument to suggest that a coarser granularity with which an explanans is described is necessarily ordinally equivalent to the accuracy with which that same explanans predicts the occurrence of a fixed explanandum. To see this, consider the following cases:

> **Case D**: Suppose that a flat in London sells for an above-average price for the UK. The probability of a London flat selling at an above-average price is .9.

Explanation D: The flat sold for an above-average price because it is in London.

Case E: It turns out that the same flat needs a full renovation. The probability that a London flat needing a full renovation sells for above the UK average is .7.

Explanation E: The flat sold for an above-average price because it is in London and needs renovation.

Case F: Necessarily, the flat is also in the South of England. The probability that a flat in the South of England sells for above the UK average is .6.

Explanation F: The flat sold for an above-average price because it is in the South of England.

It should be clear from these examples that Explanation E has a more fine-grained explanans than Explanation D, which in turn has a more fine-grained explanans than Explanation F. However, the hypothesis in Explanation D, i.e. 'the flat is in London' more accurately predicts the flat selling for an above-average rate than the more fine-grained hypothesis in Explanation E, i.e. 'the flat is in London and needs a renovation'. Here it may also be the case that Explanation D is also a better explanation of why the flat sold for a high price.

This case reveals a subtlety in my overall argument. My argument can be summarized as follows. More coarse-grained explanations are sometimes better than more fine-grained ones, and more coarse-grained explanations are sometimes such that the explanatory hypothesis predicts the explanandum evidence less accurately than a more fine-grained explanation. By contrast, the measures of explanatory power considered here necessarily track the accuracy with which an explanatory hypothesis predicts its explanandum evidence. Thus, these measures are not measures of explanatory goodness. The key point to note is that, for my argument to go through, it must only be the case that coarse-graining a hypothesis *sometimes* leads to an increase in explanatory goodness at the expense of accuracy. This need not always be the case, and in fact is not always the case, as evidenced by Explanations D-F. In those explanations, coarse-graining the hypothesis resulted in both an increase in accuracy and an increase in explanatory goodness. That this can happen does not threaten my argument against the idea that the measures considered here cannot be measures of explanatory goodness because they necessarily track the overall quality of an explanation.

Further, the case can be made that the accuracy with which an explanatory hypothesis predicts the evidence that it explains is a *pro tanto* good-making feature of an explanation. It is just that causal accuracy can be traded off against other good-making features of a causal explanation, such as a lack of complexity, such that these trade-offs can result in a better explanation overall. This echoes an argument that Andersen (2017) makes in the context of the description of a pattern of data, writing that "the more efficiently we describe a pattern, the faster we can identify whether it occurs. If there are tasks for which speed is relevant, we might prioritize efficiency of description and accept reduced accuracy" (p. 603). I argue that the same holds of explanations; reduced accuracy can sometimes be traded off for increased efficiency, for the reasons given above.

2.7 Possible Alternative Measures

In light of the limitations of the measures considered above, it is worth considering whether there are alternative formal measures of explanatory goodness that fare better. More specifically, it is worth considering whether there are measures that are not necessarily monotonically associated with the accuracy with which an explanatory hypothesis predicts the evidence that it explains. In this section, I consider two possible measures that meet this condition: one due to Ströing (2018), and my pragmatic measure, which I present in more detail in the next chapter. A common thread between these two measures is that they are functions with more arguments than those proposed by either Schupbach and Sprenger, Crupi and Tentori, or Eva and Stern. I take this to be evidence that the concept of explanatory power has more dimensions than these authors' measures can accommodate. However, in what follows, I will call attention to a drawback of Ströing's measure, namely that it achieves the positive result described above via largely *ad hoc* means.

I begin with the measure proposed by Ströing (2018). Like Sprenger and Schupbach, he begins with an algebra $\mathcal{A}_\Omega$ over the set of possible worlds Ω. The algebra contains the elements H, E, and D. The elements H and E represent an explanatory hypothesis and the evidence that it explains, as before. The new element D represents the data observed by some scientist. On Ströing's framework, 'data' is taken to be the raw form of scientific evidence, and 'evidence' is taken to be a particular pattern that may emerge from the data. Ströing argues that rather than formalizing explanatory power as a relationship between H and E, one should formalize explanatory power as a relationship between H and D, where this relationship is 'mediated' through the interpretation E of the data. To use his example, let the data proposition D represent the compiled readings from large interferometers designed to detect the existence of gravitational waves, provided that such waves exist. In this case, let the

evidential proposition E represent the interpretation of the data stream as a pattern consistent with the existence of gravitational waves, and let H represent Einstein's general relativity hypothesis predicting the existence of gravitational waves (2018, p. 436). Ströing aims to define a measure such that the explanatory power of H with respect to D, mediated through E, is high. That is, once we interpret the raw interferometer data as evidence of gravitational waves, the general relativity hypothesis H has a high degree of explanatory power.

Ströing formalizes his measure of explanatory power as follows, adapting Schupbach and Sprenger's approach to accomodate his distinction between data and evidence:

$$\varepsilon_{st}(H, E, D) = \frac{p(H|E)p(E|D) - max(p(H|\neg E), p(\neg E|D))}{p(H|E)p(E|D) + max(p(H|\neg E), p(\neg E|D))} \in [-1, 1] \qquad (2.7)$$

The term $p(E|D)$ represents the probability that a data set D will be interpreted by an agent as instantiating a pattern of evidence E, in the counterfactual situation where the data is known but not its interpretation. For example, if D represents a particular set of interferometer readings and E represents the existence of gravitational waves, $p(E|D)$ represents the probability that a scientist presented with the interferometer readings will interpret them as indicating the presence of gravitational waves. Where $p(E|D) = 1$, Ströing's measure is identical to Shupbach and Sprenger's.

Although Ströing does not specifically mention this, his measure of explanatory power allows for the possibility that a more predictively accurate hypothesis with respect to some evidence will have lower explanatory power with respect to the data from which some evidence is derived. To show this via a formal possibility result, let H_1 and H_2 be two hypotheses, and let E and D be some evidence. The following proposition is true:

Proposition 7. *The following two conditions are sufficient for it to be the case that* $\varepsilon_{st}(H_1, E, D) < \varepsilon_{st}(H_2, E, D)$*: 1)* $p(H_1|\neg E) < p(\neg E|D) < p(H_2|\neg E)$*, and 2)* $p(E|H_1) < \frac{p(E|H_2)(p(E|D)p(E)+p(\neg E|D)-P(E))}{p(H_1)p(\neg E|H_2)}$*. It is consistent with these two conditions that* $\mathfrak{B}(p(\cdot|H_1), \mathcal{E}) > \mathfrak{B}(p(\cdot|H_2), \mathcal{E})$.

Thus, unlike both Shupbach and Sprenger's measure and Crupi and Tentori's, Ströing's explanatory power measure does not track the Brier score measure of the accuracy with which conditionalizing on a hypothesis allows us to predict the value of the evidence being explained.

The significance of this result is borne out when we apply it to the contrast between Explanation A and Explanation B. Let H_A be the hypothesis that the diffusion constant of the dye in water is $1\frac{\text{mm}^2}{\text{s}}$, let H_B be the fine-grained hypothesis describing

the initial micro-state of the dye in water, let E be the proposition that a given dye particle is within a 9mm Euclidean distance from its origin at $t = 1\text{s}$, and let D be the raw data that scientists interpret in order to arrive at E (e.g., a video of the dye droplet diffusing). In the cases presented, $p(E|H_A) = .97$ and $p(E|H_B) = 1$, which leads to the result that, on the Brier score measure, H_B is more predictively accurate than H_A with respect to E. Bayes' theorem delivers the result that that $p(H_A|\neg E) = \frac{p(\neg E|H_A)p(H_A)}{p(\neg E)} = \frac{(1-p(E|H_A)p(H_A)}{p(\neg E)} = \frac{.03p(H_A)}{p(\neg E)}$, and that $p(H_B|\neg E) = \frac{0*p(H_B)}{p(\neg E)} = 0$.

Proposition 7 tells us that one way to ensure that $\varepsilon_{st}(H_B, E, D) < \varepsilon_{st}(H_A, E, D)$ is to stipulate that $0 < p(\neg E|D) < \frac{.03p(H_A)}{p(\neg E)}$ and that $1 < \frac{.97(p(E|D)p(E)+p(\neg E|D)-p(E))}{.03p(H_B)}$. The latter stipulation can be re-written as $\frac{.03}{.97} < \frac{p(E|D)p(E)+p(\neg E|D)-P(E)}{p(H_B)}$. These stipulations are implied by a set of sufficient conditions that seem to be satisfied in the case of the dye diffusing in water. Specifically, the following proposition is true:

Proposition 8. *If it is the case that* $0 < p(E) < .97$, $0 < p(H_A) < 1$, $0 < p(H_B) < \frac{97}{3}(p(E|D)p(E) + p(\neg E|D) - p(E))$, *and* $\frac{100p(E)+3p(H_A)-100}{100(p(E)-1)} < p(E|D) \leq 1$, *then it is the case that* $0 < p(\neg E|D) < \frac{.03p(H_A)}{p(\neg E)}$ *and* $\frac{.03}{.97} < \frac{p(E|D)p(E)+p(\neg E|D)-p(E)}{p(H_B)}$, *such that* $\varepsilon_{st}(H_B, E, D) < \varepsilon_{st}(H_A, E, D)$.

This gives us a set of conditions such that the diffusion constant $1\frac{\text{mm}^2}{\text{s}}$ for the dye in water is a more powerful explanation of the dye's diffusion than initial micro-state of the water. Importantly, these conditions allow for the possibility that $p(H_B) < p(H_A)$, which allows us to represent the fact that it is much more likely that the water has a particular diffusion constant for the dye than it is that the system is in a particular initial microstate. This result demonstrates that Ströing's measure can successfully track our judgments about the better explanation in some scientific cases where both Schupbach and Sprenger and Crupi and Tentori's measures fail. Although the Bayes nets formalism does not contain any tools for representing the distinction between data and evidence, I take it that the Bayes nets framework can be suitably augmented so that a similar result could be delivered for a version of Ströing's measure that is adapted for the Bayes nets context. The results above suggest that on such an augmented Bayes nets framework, Ströing's measure could fare better than Eva and Stern's in tracking our judgments as to the better explanation in the case considered.

However, the conditions that are used to generate these positive results for Ströing's measure are decidedly *ad hoc*. For instance, it is not at all obvious why the hypothesis H_A should count as a better explanation of the evidence E than the hypothesis H_B in virtue of satisfying the sufficient conditions in Proposition 8. Thus, while Ströing's

measure may solve a mathematical problem that besets the other measures considered here, it does not resolve the mismatch between the mathematical features of these measures and actual scientific practice. Where scientists prefer less predictively accurate, more coarse-grained hypotheses for explanatory purposes, this preference is not due to the fact that the mathematical conditions identified above are satisfied. Thus, I do not take Ströing's measure to be an appealing solution to the issues raised here for probabilistic measures of explanatory power.

The following chapter takes an entirely different approach to measuring the goodness of an explanation, by focusing on the pragmatic value of the information provided by an explanatory hypothesis, for an agent with a prudential interest in predicting the evidence that the hypothesis explains. To illustrate via an example, suppose that an agent is betting on whether a given particle of the dye dropped in water will be within a 9mm Euclidean distance from its origin at $t = 1\text{s}$. Suppose further that such an agent will bet the same way whether they learn the diffusion constant of the dye in water, or the micro-state of the water. From a behavioral perspective, there is no scenario in which learning the more coarse-grained information will lead them to bet one way, and learning the more fine-grained information will lead them to bet another way. This implies that the agent would pay just as much to learn the diffusion constant as the micro-state, where the specific amount that the agent would pay for either piece of information is determined by the payouts of the various possible bets. This is possible even when the micro-state is a better predictor of the diffusion of the dye in water than the diffusion constant. Assuming further that agents prefer simpler explanations, these pragmatic considerations all weigh in favor of the explanations citing the diffusion constant, and against the explanation citing the micro-state. Thus, in the next chapter, I provide a mathematically precise, pragmatically motivated procedure for determining when we might prefer a less predictively accurate explanatory hypothesis over a more predictively accurate hypothesis.

It is worth noting what both of these alternative measures have in common. Both measures have additional inputs besides the probabilistic relationship between an explanatory hypothesis and the evidence it explains. Ströing's measure takes as an input the probabilistic relationship between evidence and data, while my measure takes as an additional input an agent's utility function over different possible states of the evidence under different possible bets. That both measures are able to come apart from the accuracy with which an explanatory hypothesis predicts the evidence that it explains suggests that Schupbach and Sprenger, Crupi and Tentori, and Eva and Stern's measures are all constrained by their comparatively low number of inputs.

2.8 Conclusion

This chapter has shown that many attempts to measure the power with which a hypothesis explains some evidence have in fact measured the accuracy with which a hypothesis predicts the evidence that it explains. In scientific practice, we sometimes prefer explanations that trade off predictive accuracy for various other explanatory virtues. As such, these attempts at measuring explanatory power are not suitable measures of the overall goodness of an explanation, despite some claims to the contrary. Ströing's measure manages to circumvent this issue by taking additional arguments as inputs beyond the probabilistic relationships between hypothesis and evidence, but this measure has other drawbacks and limitations. I take these results to show that measuring the overall quality of an explanation may be a more complicated task than has previously been recognized.

As discussed above, in the next chapter I will provide a positive argument showing one way to select the optimal level of granularity for a causal explanation. Specifically, I will outline my proposed methodology for measuring the pragmatic value of a causal explanation given at a particular level of granularity, and show that this measure of pragmatic value recovers scientific judgments regarding the optimal level of granularity for describing a given event.

2.9 Appendix

2.9.1 Proof of Proposition 4

Proof. We will show that $\varepsilon_{ss}(H, E)$ is a strictly increasing function of $p(E|H)$, and that $\mathfrak{B}(p(\cdot|H), \mathcal{E})$ is a strictly decreasing function of $p(E|H)$. Let us begin with $\varepsilon_{ss}(H, E)$. Using Bayes' theorem and some algebra, we re-write this equation as follows.

$$\varepsilon_{ss}(H, E) = \frac{p(E|H)(1 - p(E)) - (1 - p(E|H))p(E)}{p(E|H)(1 - p(E)) + (1 - p(E|H))p(E)} \tag{2.8}$$

We treat $p(E)$ as a constant, take the partial derivative of $\varepsilon_{ss}(H, E)$ with respect to $p(E|H)$, and simplify it as follows:

$$\frac{\partial \varepsilon}{\partial p(E|H)} = \frac{2p(E) - 2p(E)^2}{(p(E|H) + p(E) - 2p(E|H)p(E))^2} \tag{2.9}$$

Since $p(E)$ is strictly between 0 and 1, the numerator and denominator of the right-hand side of (2.9) are both positive. Thus, the partial derivative is strictly

increasing, meaning that for a fixed $p(E)$, $\varepsilon_{ss}(H_1, E) > \varepsilon_{ss}(H_2, E)$ if and only if $p(E|H_1) > p(E|H_2)$.

Next, we consider the Brier score, which is written as follows.

$$\mathfrak{B}(p(\cdot|H), \mathcal{E}) = \frac{1}{n+1}\left(\left(p(E|H) - 1\right)^2 + \sum_{i=1}^{n}\left(p(I_i|H) - 0\right)^2\right) \tag{2.10}$$

This can be re-written as follows.

$$\mathfrak{B}(p(\cdot|H), \mathcal{E}) = \frac{1}{n+1}\left(p(E|H) - 1\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(p(I_i|H) - 0\right)^2 \tag{2.11}$$

Clearly, $\frac{1}{n+1}(p(E|H)-1)^2 \propto \frac{1}{p(E|H)}$. Further, under the stipulation that $\sum_{i=1}^{n} p(I_i|H_1) - p(I_i|H_2) = p(E|H_1) - p(E|H_2)$, $\frac{1}{n+1}\sum_{i=1}^{n}\left(p(I_i|H)-0\right)^2 \propto \frac{1}{p(E|H)}$. Thus, $\mathfrak{B}(p(\cdot|H), \mathcal{E})$ is strictly increasing at $p(E|H)$ approaches 1.

Thus, $\mathfrak{B}(p(\cdot|H_1), \mathcal{E}) < \mathfrak{B}(p(\cdot|H_2), \mathcal{E})$ if and only if $p(E|H_1) > p(E|H_2)$. Since $\varepsilon_{ss}(H_1, E) > \varepsilon_{ss}(H_2, E)$ if and only if $p(E|H_1) > p(E|H_2)$ and $\mathfrak{B}(p(\cdot|H_1), \mathcal{E}) < \mathfrak{B}(p(\cdot|H_2), \mathcal{E})$ if and only if $p(E|H_1) > p(E|H_2)$, it follows that $\varepsilon_{ss}(H_1, \mathcal{E}) > \varepsilon_{ss}(H_2, \mathcal{E})$ if and only if $\mathfrak{B}(p(\cdot|H_1), \mathcal{E}) < \mathfrak{B}(p(\cdot|H_2), \mathcal{E})$. □

2.9.2 Proof of Proposition 5

Proof. Clearly, $\varepsilon_{ct}(H, E)$ is an increasing function of $p(E|H)$ when $p(E|H) \geq p(E)$ and when $p(E|H) < p(E)$. Further, when $p(E|H) > p(E)$, $\varepsilon_{ct}(H, E) > 0$, when $p(E|H) = p(E)$, $\varepsilon_{ct}(H, E) = 0$, and when $p(E|H) < p(E)$, $\varepsilon_{ct}(H, E) < 0$. Thus, $\varepsilon_{ct}(H, E)$ is an increasing function of $p(E|H)$ for all possible conditions. Since we have established in the previous proof that $\mathfrak{B}(p(\cdot|H), \mathcal{E})$ is strictly decreasing as $p(E|H)$ approaches 1, the proposition is true. □

2.9.3 Proof of Proposition 6

Proof. Since $\varepsilon_{es}(h_j, e_i, \mathcal{N})$ has the same mathematical form as $\varepsilon_{ss}(H, E)$, the first steps from the proof of Proposition 4 can be repeated to show that, for a fixed $p(E)$, $\varepsilon_{es}(h_j, e_i, \mathcal{N})$ is a strictly increasing function of $p(e|do(h_j))$. This implies that $\varepsilon_{es}(h_j^1, e_i, \mathcal{N}) > \varepsilon_{es}(h_l^2, e_i, \mathcal{N})$ if and only if $p(e_i|do(h_j^1)) > p(e_i|do(h_l^2))$.

Next, we consider the causal Brier score $B_c(p(\cdot|do(h_j), E, \mathcal{N})$, which is written out as follows.

$$\mathfrak{B}_c(p(\cdot|do(h_j)), E, \mathcal{N}) = \frac{1}{n}(p(e_i|do(h_j)) - 1)^2 + \frac{1}{n}\sum_{k=1|i\neq k}^{n}(p(e_k|do(h_j)) - 0)^2 \tag{2.12}$$

Clearly, $\frac{1}{n}(p(e_i|do(h_j)) - 1)^2 \propto \frac{1}{p(e_i|do(h_j))}$. Further, since any decrease in $p(e_i|do(h_j))$ will lead to an increase in one or more probabilities $p(e_k|do(h_j))$, where $e_i \neq e_k$, $\sum_{k=1|i\neq k}^{n}\frac{1}{n}(p(e_k|do(h_j)) - 0)^2) \propto \frac{1}{p(e_i|do(h_j))}$. Thus, $\mathfrak{B}_c(p(\cdot|do(h_j)), E, \mathcal{N})$ is strictly decreasing as $p(e_i|do(h_j))$ approaches 1. This entails that $\mathfrak{B}_c(p(\cdot|do(h_j^1)), E, \mathcal{N}) < \mathfrak{B}_c(p(\cdot|do(h_l^2)), E, \mathcal{N})$ iff $p(e_i|do(h_j^1)) > p(e_i|do(h_l^2))$. Since $\varepsilon_{es}(h_j^1, e_i, \mathcal{N}) > \varepsilon_{es}(h_l^2, e_i, \mathcal{N})$ iff $p(e_i|do(h_j^1)) > p(e_i|do(h_l^2))$ and since $\mathfrak{B}_c(p(\cdot|do(h_j^1)), E, \mathcal{N}) < \mathfrak{B}_c(p(\cdot|do(h_l^2)), E, \mathcal{N})$, iff $p(e_i|do(h_j^1)) > p(e_i|do(h_l^2))$, one can conclude that $\varepsilon_{es}(h_j^1, e_i, \mathcal{N}) > \varepsilon_{es}(h_l^2, e_i, \mathcal{N})$ iff $\mathfrak{B}_c(p(\cdot|do(h_j^1)), E, \mathcal{N}) < \mathfrak{B}_c(p(\cdot|do(h_l^2)), E, \mathcal{N})$. □

2.9.4 Proof of Proposition 7

Proof. We need to identify a set of sufficient conditions for it to be the case that $\varepsilon_{st}(H_1, E, D) < \varepsilon_{st}(H_2, E, D)$. We can expand this inequality as follows:

$$\frac{p(H_1|E)p(E|D) - max(p(H_1|\neg E), p(\neg E|D))}{p(H_1|E)p(E|D) + max(p(H_1|\neg E), p(\neg E|D))} < \frac{p(H_2|E)p(E|D) - max(p(H_2|\neg E), p(\neg E|D))}{p(H_2|E)p(E|D) + max(p(H_2|\neg E), p(\neg E|D))} \tag{2.13}$$

If we stipulate that $p(H_1|\neg E) < p(\neg E|D) < p(H_2|\neg E)$, then we can reduce (2.13) as follows:

$$\frac{p(H_1|E)p(E|D) - p(\neg E|D)}{p(H_1|E)p(E|D) + p(\neg E|D)} < \frac{p(H_2|E)p(E|D) - p(H_2|\neg E)}{p(H_2|E)p(E|D) + p(H_2|\neg E)} \tag{2.14}$$

With some algebra, one can show that, under the assumption that all probabilities are in the interval $(0, 1)$, (2.14) is true if $p(E|H_1) < \frac{p(E|H_2)(p(E|D)p(E)+p(\neg E|D)-P(E))}{p(H_1)p(\neg E|H_2)}$. Thus, if $p(H_1|\neg E) < p(\neg E|D) < p(H_2|\neg E)$ and it is the case that $p(E|H_1) < \frac{p(E|H_2)(p(E|D)p(E)+p(\neg E|D)-P(E))}{p(H_1)p(\neg E|H_2)}$, then $\varepsilon_{st}(H_1, E, D) < \varepsilon_{st}(H_2, E, D)$.

It is consistent with these two conditions that $\mathfrak{B}(p(\cdot|H_1), \mathcal{E}) > \mathfrak{B}(p(\cdot|H_2), \mathcal{E})$. To see why, let it be the case that $p(E|H_1) < p(E|H_2)$, so that $\mathfrak{B}(p(\cdot|H_1), \mathcal{E}) > \mathfrak{B}(p(\cdot|H_2), \mathcal{E})$. Let it also be the case that $p(H_1|\neg E) = .4$, $p(\neg E|D) = .5$, and $p(H_2|\neg E) = .6$, so that the first condition derived above holds. Suppose further that $p(E) = .5$. One can check that for all $p(H_1) < \frac{9}{20}$, there are values of $p(H_2)$ such that

$p(E|H_1) < \frac{p(E|H_2)(p(E|D)p(E)+p(\neg E|D)-P(E))}{p(H_1)p(\neg E|H_2)}$, so that both conditions derived above are satisfied. □

2.9.5 Proof of Proposition 8

Proof. In the specific case under consideration, we know that $p(E|H_A) = .97$ and $p(E|H_B) = 1$. Proposition 7 tells us that if $0 < p(\neg E|D) < \frac{.03p(H_A)}{p(\neg E)}$ and $\frac{.03}{.97} < \frac{p(E|D)p(E)+p(\neg E|D)-p(E)}{p(H_B)}$, then $\varepsilon_{st}(H_B, E, D) < \varepsilon_{st}(H_A, E, D)$. What we need to do is determine values for the probabilities $p(E)$, $p(H_A)$, $p(H_B)$, and $p(E|D)$ such that these inequalities are true. That is, we need to solve the following system of inequalities:

$$0 < p(\neg E|D) < \frac{.03p(H_A)}{p(\neg E)} \tag{2.15}$$

$$\frac{.03}{.97} < \frac{p(E|D)p(E) + p(\neg E|D) - p(E)}{p(H_B)} \tag{2.16}$$

$$0 < p(E) < 1 \tag{2.17}$$

$$0 < p(H_A) < 1 \tag{2.18}$$

$$0 < p(H_B) < 1 \tag{2.19}$$

$$0 < p(E|D) < 1 \tag{2.20}$$

One solution to this system of inequalities is $0 < p(E) < .97$, $0 < p(H_A) < 1$, $0 < p(H_B) < \frac{97}{3}(p(E|D)p(E) + p(\neg E|D) - p(E))$, and $\frac{100p(E)+3p(H_A)-100}{100(p(E)-1)} < p(E|D) \leq 1$. This proves the proposition. □

3 On The Explanatory Depth and Pragmatic Value of Coarse-Grained, Probabilistic, Causal Explanations

3.1 Introduction

As discussed in the preceding chapters, the special sciences frequently make probabilistic generalizations. To take a classic example, a patient who habitually smokes cigarettes is said to have a certain probability of developing lung cancer. These generalizations often facilitate causal explanations. In the case of smoking and lung cancer, if we ask *why* a given smoker develops lung cancer, we can explain their affliction by saying that smoking was the likely cause. Similarly, if we observe a population in which more smokers develop lung cancer than non-smokers, then we can say that the higher incidence of lung cancer among smokers is causally explained by the higher probability with which smokers get lung cancer.

The previous chapters have established that in each of these cases, we could also tell a more complicated story. Instead of explaining a patient's lung cancer by citing the fact that they smoked, we could attempt to cite a long series of facts about individual strands of burnt tobacco producing carcinogenic particles that entered the smoker's lungs at different spatiotemporal points, eventually causing them to develop lung cancer. Clearly, there are pragmatic reasons for avoiding this kind of reduction. The level of detail required for the second kind of explanation would be incredibly cumbersome, and clearly not suited for the task at hand. In addition, many recent authors in philosophy of science, including Strevens (2008), Woodward (2010, 2016), Weslake (2010, 2013), Weatherson (2012), Franklin-Hall (2016) and Clarke (2017), have argued that coarse-grained explanations are to be preferred on explanatory as well as pragmatic grounds.[1] The central idea that each of these authors express is that there is a sense in which some coarse-grained causal generalizations provide *deeper* explanations than their more fine-grained counterparts. In the previous chapter, I described how Weslake defines explanatory

[1] Historically, similar positions have been held by Putnam (1979), Kitcher (1981), Garfinkel (1981), Jackson and Pettit (1992), Sklar (1993), and Batterman (2001).

depth as "a measure in terms of which explanations can be assessed according to their explanatory value" (2010, 273). I continue to adopt this definition here; explanatory depth refers to the relative value of an explanation *qua* explanation. Thus, 'explanatory depth' refers to the same property of explanations as 'explanatory goodness'.

Most of these authors consider, and accept to varying degrees, the thesis that higher-level causal explanations are deeper than lower-level explanations when and because the causes cited are *proportional* to their effects. The meaning of proportionality has been discussed in Chapter 1 and will be discussed again in this chapter, but the basic idea is as follows. When a cause and an effect are both represented by a variable taking some value, the cause is proportional to the effect to the extent that it is true that if the causal variable had taken a different value, then the effect variable would also have taken a different value. Notably, Weslake (2010, 2013) draws a link between the relative proportionality and the relative *abstractness* of an explanation, where abstractness is the ability of an explanatory causal model to be deployed in a wide range of scenarios. Abstractness, he argues, provides a crucial link between proportionality and explanatory depth; proportional explanations are deeper than less proportional ones because they are more abstract.

However, the linkage between proportionality, abstractness, and explanatory depth has largely been developed in deterministic contexts. In what follows, I argue that in probabilistic contexts, it can be the case that a more proportional explanation is less abstract than a less proportional explanation of the same explanandum. In some of these cases, the more proportional explanation seems to be deeper, and in others, the more abstract explanation seems to be deeper. Thus, probabilistic cases of causal explanation muddy the waters when it comes to the relationship between proportionality, abstractness and depth. In these cases, I claim that we need to consider the pragmatic value of an explanatory model in order to make sense of its relative explanatory depth. Thus, I reject the strict bifurcation between pragmatics and explanatory depth, and conclude that pragmatics plays an indispensable role in assessing the relative depth of competing causal explanations.

The plan for this chapter is as follows. In Section 3.2, I provide some background on probabilistic causal explanation and the formal notion of the granularity of explanation. In Section 3.3, I explicate the nature of the relationship between proportionality, abstractness, and explanatory depth, as it is developed in deterministic contexts. In Section 3.4, I turn to probabilistic contexts. I introduce a probabilistic account of proportionality put forward by Pocheville et al. (2017), and use it to show how proportionality and explanatory depth can come apart in cases where subtle differences in the nature of a cause lead to slight differences in the probability of its effect. I then show how abstractness and explanatory depth can also come apart in

probabilistic cases. In Section 3.5, I introduce my proposed measure of the pragmatic value of a causal model, using the decision-theoretic concept of information pricing. I use this measure to argue that the pragmatic value of a causal explanation allows us to determine the optimal level of granularity for that explanation in a way that proportionality and abstractness are unable to do. In Section 3.6, I address some possible objections to my argument. In Section 3.7, I offer concluding remarks.

3.2 Background

Throughout this chapter, I take as a background assumption the definition of a causal explanation that is outlined in Chapter 1, sub-section 1.2.2. Recall that according to this definition, one event causally explains another if an intervention bringing about the explanans event increases the probability of the explanandum event, according to a Bayes net that is representationally adequate with respect to the system in which the events occur. Additionally, throughout this chapter, I make reference to coarsenings of variables in a probabilistic causal model or Bayes net. Recall that a variable X is a *coarsening* of another variable X' if and only if the value set of X is a *quotient set* of the value set of X', and the value set of X has less elements than the value set of X'. I assume that these all variables used in this chapter are elements of an "extended" Bayes net that contains intervention variables, as defined in Section 1.2.3.

3.3 Depth, Proportionality, and Abstractness

The same event can be given a different causal explanation in virtue of changing the granularity of the causal variable. To illustrate, consider a variable X, which takes the value x_1 if a person smokes and x_2 if they do not. Next, we can introduce a variable L which takes the value l_1 if a person has lung cancer, and l_2 if they do not. Interventions that change the value of X also change the probability distribution over L; if we intervene so that a person becomes a smoker, then they are more likely to develop lung cancer, and if we intervene so that a person does not become a smoker, then they are less likely to develop lung cancer. Thus, on my model of causal explanation, the fact that a person smokes causally explains why they get lung cancer.

Next, let Y be a variable that takes the value y_1 if a person smokes Marlboros, y_2 if they smoke another brand, and y_3 if they do not smoke. Here, there are *at least some* interventions on the value of Y that change the probability distribution over L, i.e. those interventions that change the value of Y from either y_1 to y_3 or y_2 to

y_3, or vice-versa. It may also be the case that all interventions on Y change the probability distribution over L, if Marlboros are more or less carcinogenic than other brands, but this is not necessary for a causal relationship; all that is required is that at least some interventions on the causal variable lead to changes in the probability distribution over the effect variable. If we make the further plausible assumption that smoking Marlboros raises the probability of developing lung cancer relative to the prior probability of lung cancer, then we can say that if a patient who smokes Marlboros develops lung cancer, then their smoking Marlboros causally explains their lung cancer.

In a sense, this is exactly the result that we want. It is true that the fact that a patient smokes Marlboros causally explains why they develop lung cancer. However, there is also a sense in which the explanation is suboptimal. Instead of explaining the patient's lung cancer by citing the fact that they smoked Marlboros, we could say that the patient developed lung cancer because they smoked. This second, more coarse-grained explanation seems preferable. After all, the brand of cigarettes that a person smokes does not matter much, if at all, to their chances of getting lung cancer; the relevant detail here is that the patient smokes. Providing a rationalization for this kind of intuition is one goal of the literature on the explanatory depth of causal explanations.

As discussed in Chapter 1, a popular idea in the literature on higher-level causal explanation is that more coarse-grained causal explanations can be chosen over more fine-grained explanations of the same causal phenomenon on the basis that the more coarse-grained explanation is more *proportional* than the more fine-grained one. This idea appears in Yablo (1992), and is endorsed to varying degrees by Craver (2007), Woodward (2010), Malaterre (2011), Weslake (2013), and Stegmann (2014). Note that whereas in Chapter 1, proportionality was considered as a putatively necessary feature of a causal explanation, here proportionality is considered as a putatively *good-making* feature of a causal explanation. Woodward (2010) provides perhaps the most comprehensive definition of proportionality as a property of the relationship between cause and effect. For Woodward, proportionality is a property of causal relationships that comes in degrees. A causal variable stands in a proportional relationship to a given effect variable to the extent that the causal variable is described so that: (a) each possible intervention changing the value of the causal variable determines the value of the effect variable, and (b) all possible interventions changing the value of the causal variable correspond to changes in the value of the effect variable (Woodward 2010, p. 296).

We can use a classic example from Yablo (1992) to illustrate Woodward's understanding of proportionality. Suppose that we have trained a pigeon to peck at only red targets, to the point that the pigeon will peck at any red target, and refrain from

pecking at targets of any other color, as a matter of certainty. Now suppose that we present the pigeon with a scarlet target, leading it to peck. We can say that the pigeon pecked at the target either because the target was scarlet or because the target was red. Clearly, saying that the pigeon pecked because the target was scarlet fails to fully satisfy proportionality. Specifically, there is a failure to satisfy condition (b). We can intervene on the target to change it from scarlet to crimson, and the pigeon will still peck. By contrast, any intervention that changes the color of the target from red to some other color will make it the case that the pigeon does not peck. So the explanation 'the pigeon pecked because the target was red' perfectly satisfies (a) and (b), whereas 'the pigeon pecked because the target was scarlet' does not.

However, this does not answer the question of why we regard proportional explanations as being deeper than less proportional ones. Weslake (2010) argues that part of what is at work here is that we tend to regard more *abstract* explanations as being deeper than less abstract ones. For Weslake, when we compare two explanations, the more abstract explanation is one that we can use in a wider variety of cases, without changing the values of the variables used. Every Marlboro smoker is also a smoker, so the explanation 'smoking caused the patient's lung cancer' will be applicable in all cases in which the explanation 'smoking Marlboros caused the patient's lung cancer' is applicable. However, the reverse obviously does not hold; the explanation 'smoking Marlboros caused the patient's lung cancer' does not apply in all the cases in which 'smoking caused the patient's lung cancer' does.

Weslake (2013) argues further that there is a clear connection between the abstractness of a causal explanation and its proportionality. If we consider the pigeon example, it is clear that the explanation 'the pigeon pecked because the target was scarlet' is applicable to less cases than 'the pigeon pecked because the target was red', and that in any case in which the former applies, the latter also applies. Indeed, in deterministic cases a perfectly proportional causal explanation of a given explanandum will generally be more abstract than a less proportional explanation of the same explanandum. If all red targets cause the pigeon to peck, and all non-red targets cause the pigeon not to peck, it follows that any red target that has some other property A will also cause the pigeon to peck, and any non-red target that has some other property B will cause the pigeon not to peck. Thus, a more proportional explanation is applicable wherever a less proportional one is also applicable. However, it follows from the truth of the more proportional explanation that the pigeon will also peck when presented with a red target that does not have property A, and not peck when presented with a non-red target that does not have property B. The explanation 'the pigeon pecked because it was presented with a red target with property A' will not apply in these cases, but the more proportional explanation

'the pigeon pecked because the target was red' will. Thus, proportionality tracks abstractness in deterministic cases.

Moving further, there is also a strong intuitive connection between abstraction and explanatory depth. I take it that the force of this intuition lies in the thought, most famously expressed by Kitcher (1981), that explanation in general involves the subsumption of specific events under more general patterns, in this case patterns of causation. If such subsumption is part of the purpose of explanation, then it makes sense that we would prefer to give explanations that are able to subsume more events under a given type-level causal pattern. This preference for abstraction is linked closely to what Ylikoski and Kuorikoski call the "cognitive salience" of a given explanation. They write that "the kinds of inferences possible for limited cognitive systems such as humans directly effect what can be explained and understood by such cognitive systems" (2010, p. 214). Since more abstract explanations tend to use more coarse-grained variables, they allow us to use "the same words [to refer] to similarly structured forms of causal interaction in different fields," thereby reducing the cognitive burden associated with causal explanation in many cases (p. 214).

Weslake argues that an emphasis on abstraction as a good-making feature of an explanation stands in contrast with Hitchcock and Woodward's (2003) claim that causal explanations are deeper when and because they are situated within an explanatory model that is more "invariant", where a more invariant model is one that answers more what-if-things-had-been-different questions about the target system. To use Weslake's example, the ideal gas law can be used to explain the change in the pressure of a gas with a fixed temperature and changing volume. That is, we can use the ideal gas law to answer questions about what would happen under various changes in the volume of the gas' container. However, a model containing a variable representing the microphysical state of the gas, rather than its volume, answers more counterfactual questions than the more abstract, coarse-grained model. For Weslake, this shows the limitations of invariance as a dimension of explanatory depth; highly fine-grained explanations seem to have greater or equal invariance than coarse-grained explanations in cases where a coarse-grained explanation seems more appropriate. I concur with Weslake's argument here, although see Woodward (2018) for a response.

3.4 The Probabilistic Context

In probabilistic contexts, the picture presented above becomes considerably more complicated. So we will need to translate the concepts of proportionality and abstractness to accomodate probabilistic relationships between variables. In what

follows, I will put forward what I take to be reasonable probabilistic extensions of the concepts of proportionality and abstractness. I will show that there are some cases in which a *less* proportional causal explanation seems to be deeper than a more proportional one. Similarly, I will show that in other cases, a *less* abstract explanation seems to be deeper than a more abstract competitor. Thus, I conclude that the tight relationship between proportionality, abstractness and explanatory depth comes apart in probabilistic contexts.

Let us begin by defining a probabilistic measure of proportionality. Clearly, the understanding of proportionality articulated above is not suited to probabilistic relationships, since it measures the proportionality of a causal relationship according to the extent to which the various states of the cause determine the state of the effect. So we will need a new measure for assessing the comparative proportionality of causal relationships where the relationship between cause and effect is probabilistic. One attempt at this task is put forward by Pocheville et al. (2017), who understand proportionality as follows. If refining the causal variable makes no difference to the conditional probability distribution over some effect variable, then the relationship between the coarse-grained causal variable and the effect variable is more proportional than the relationship between the fine-grained causal variable and the effect variable. However, if refining the causal variable makes *any difference at all* to the conditional probability distribution over the effect variable, then the relationship between the fine-grained causal variable and the effect variable is more proportional than the relationship between the coarse-grained causal variable and the effect variable.

To see why this translation of proportionality into the probabilistic context is warranted, note that, in general, the concept of proportionality aims to capture the salient *difference-making* relationship between cause and effect (see List and Menzies 2009). In the pigeon case, when we say that the pigeon pecked at the target because it was scarlet, we fail to identify what is really making the difference with respect to whether the pigeon pecks. It is the target's redness, rather than its being any shade of red, that makes the difference. A perfectly proportional causal explanation is one that conveys all and only the difference-makers with respect to some effect. In probabilistic cases, it seems that we have no choice but to interpret this difference-making constraint probabilistically, and say that the proportional explanation is the one that identifies every factor that makes a difference to the probability distribution over the effect, and only those factors.[2]

[2]Note that in Pocheville et al.'s paper, they aim to define proportionality and other aspects of causal relationships using information theory, as formulated by Shannon and Weaver (1949). On their view, if the mutual information between an intervention on a causal variable and its effect is greater than that between another cause-effect pair, then the first relationship is more proportional. If two relationships have equal mutual information, the more proportional relationship is that with minimal entropy. This criterion ensures that any refinement of the causal variable that makes any difference to

We can show that Pocheville et al.'s criterion gets the right results in some intuitive cases. Consider the following case. Suppose that the probability of a smoker developing lung cancer is .1, and the probability of a non-smoker developing lung cancer is .01. Suppose that the brand of cigarette smoked does not make a difference to the patient's likelihood of lung cancer; both Marlboro smokers and non-Marlboro smokers have a .1 probability of developing lung cancer. Pocheville et al.'s approach renders the verdict that 'smoking caused the patient's lung cancer' is a more proportional explanation than 'smoking Marlboros caused the patient's lung cancer'. This fits with common sense as well; in cases where the brand of cigarette smoked makes no difference to a person's probability of developing lung cancer, the more proportional, more abstract, and deeper explanation is one that eschews any mention of cigarette branding.

So far, so good. However, once we let go of the constraint that refining the causal variable makes no difference to the probability of the effect, proportionality seems to come apart from explanatory depth. Suppose that smoking Marlboros results in a .1001 probability of lung cancer, and smoking other brands results in a .0999 probability of lung cancer. As before, smoking in general results in a .1 probability of lung cancer, and non-smoking results in a .01 probability of lung cancer. It follows from Pocheville et al.'s approach that the explanation 'smoking Marlboros caused the patient's lung cancer' is more proportional than the explanation 'smoking caused the patient's lung cancer'. However, I take it that the latter explanation is deeper than the former. Certainly, in the actual practice of science it would seem overly specific to cite the brand of cigarettes that a person smokes when explaining why they developed lung cancer, simply because the person chose a very slightly more carcinogenic brand. What seems to matter here for explanatory purposes is that there is an important causal trend—smoking causes lung cancer—that a given case can be subsumed under. Thus, proportionality and explanatory depth come apart in this case.

Next, we can extend the notion of abstractness into the probabilistic setting. Recall that causal explanation is defined above as probability-raising under intervention; $C = c_j$ explains $E = e_i$ if and only if an intervention setting the value of C to c_j raises the probability that $E = e_i$. Next, recall that the most abstract explanation of a given explanandum event is one that subsumes the widest class of specific events under a given causal relation. Thus, 'the patient's smoking caused their lung cancer' is a more abstract explanation than 'the patient's smoking Marlboros caused their lung cancer'. More generally, we can say that the most abstract explanation of a given explanandum is one that subsumes under one broad category all the ways in which the probability of the explanandum event could be raised via intervention.

the conditional probability distribution over the effect variable will yield a more proportional causal relationship.

Typically, we think that abstractness in explanation tends to go hand-in-hand with explanatory depth; explanations that are not optimally abstract are usually overly detailed and therefore suboptimal.

However, there are cases that put pressure on the notion that explanations are deeper when and because they are more abstract. Consider a case where a person suffering from some bacterial disease receives an antibiotic containing penicillin and recovers. Suppose that in general, antibiotics containing penicillin have a .95 probability of curing patients with this disease. However, suppose that doctors could also prescribe the patient an antibiotic that does not contain penicillin, and that in these circumstances the patient would have recovered with probability .7. Without any antibiotics, the recovery rate is only .1. For some assignments of probabilities to each causal intervention, administering either antibiotic raises the patient's probability of recovery, as compared to the marginal probability of recovery obtained by taking a weighted average of the conditional probability of recovery under each possible intervention. Thus, a more abstract rival to the explanation 'an antibiotic containing penicillin caused the patient to recover' would be the explanation 'an antibiotic caused the patient to recover'. However, it is not clear to me that in this case, the more abstract explanation is the better or deeper one. It seems to matter here that the patient's recovery was aided by penicillin, and leaving this fact out seems to miss something important about the patient's recovery. To use Franklin-Hall's phrase, the more abstract explanation "overshoots" the optimal level of explanatory granularity (2016, p. 570). Thus, in some probabilistic cases, abstraction does not track explanatory depth.

In light of these arguments, we will need to look for another way of determining the optimal level of granularity for explanation in probabilistic cases. In the next section, I will introduce my strategy for measuring the pragmatic value of a causal model. I will then argue that by considering the pragmatic value of a causal model, we can arrive at a compelling justification of our intuitions regarding explanatory depth in the cases described above.

3.5 The Pragmatic Value of a Causal Model

It is clear from the arguments in the previous section that in probabilistic contexts, an explanation can be maximally proportional but overly detailed, because the differences that various refinements of the causal variable make to the probability distribution over the effect variable are too small. On the other hand, an explanation can be too abstract; sometimes details that make a difference to the probability distribution add explanatory depth, even if we can pick out the class of probability-

raising factors for some explanandum without using these details. Weatherson (2012) calls this difficulty in finding the best level of detail for a given explanation the "Goldilocks Problem"; we need a level of granularity for our causal explanations that is not too detailed, not too abstract, but just right.

Franklin-Hall writes that "making sense of the explanatory—not merely the practical—superiority of high-level explanations in a physical world has been a kind of Holy Grail in the philosophy of science, long sought but never found" (2016, p. 555). My own thought is that a distinctly non-pragmatic dimension to explanatory depth may not exist. Indeed, when we rationalize the intuition that a given detail ought to be included or not included in the deepest explanation of some event, it seems that we invariably revert to pragmatic language. If asked why we do not usually include the brand of cigarette smoked in an explanation of why a patient developed lung cancer, even when a patient smokes a slightly more carcinogenic cigarette, there is a strong urge to say that the slight difference in brands just does not *matter* enough to be worth documenting in an explanation. By contrast, in the antibiotics case it *does* seem to matter that penicillin was administered to the patient. This emphasis on the degree to which a given detail matters already shifts us into a consideration of the pragmatic value of an explanation. So it makes sense that we would appeal to pragmatics in assessing the overall superiority of a higher-level causal explanation as compared to a lower-level alternative.

There is a useful analogy here to the epistemic standard required for conviction in a criminal trial. Juries are asked to determine whether the defendant is guilty "beyond a reasonable doubt". Of course, the nature of this standard is vague. If by 'beyond a reasonable doubt' we mean absolute certainty, then few if any defendants will ever be found guilty. So it must be that a juror's credence in the defendant's guilt must rise above some threshold probability. This generates the problem of where to set the threshold. It is at this point that pragmatics will begin to play a role. If the defendant is facing life in prison, then the threshold will have to be set very high, to account for the negative consequences of making a mistake. If the possible sentence is much less serious, then the jury's standard will likely drop as well. I take it that a similar dynamic is at play in the pragmatic evaluation of explanatory models. There is some threshold for the degree of probabilistic difference-making between possible values of a causal variable such that those values should be included in our model. However, setting the threshold at 'any difference at all' is unrealistic and unintuitive; it would yield hopelessly complicated explanatory models. What we need to do is consider the pragmatic consequences of including or not including the detail in an explanatory model, and set our threshold for explanatorily relevant difference-making accordingly.

	Lung Cancer	No Lung Cancer
Approve	-100	10
Deny	0	0

Tab. 3.1: Utility Matrix For Life Insurance Decision

Thus, I will seek to develop a distinctly pragmatic criterion for determining the best level of description for a given probabilistic explanation. My approach can be summarized as follows. When a scientist or ordinary person explains a phenomenon causally, there is a sense in which they make a decision. They take a particular cause of some effect and choose to embed it within a given causal model, as one value of a causal variable that may be defined at many different levels of granularity. I argue that in making this decision, the agent should choose the most coarse-grained causal variable that does not elide any information that is of value to them. To justify this preference for coarse-grained causal explanations, recall both Weslake's (2010) and Ylikoski and Kuorikoski's (2010) insistence that it is a virtue of some explanatory models that we can use the same names to refer to similar causal factors, thereby providing a unified understanding of why some phenomenon occurred. My criterion takes this insistence on coarse-graining seriously, arguing that if there is no pragmatic cost to an agent from coarsening their description of a causal factor, then that agent should coarse-grain.

My proposal measures the value of information using the decision-theoretic concept of the fair price of causal information (FPCI), which has its roots in Blackwell (1951), Savage (1954) and Good (1967), and is also discussed in Resnik (1987). To illustrate the concept, consider the following example. Suppose that an analyst has to decide whether to approve or deny an individual for life insurance. In an obvious oversimplification, suppose that only one fact—whether or not the potential customer develops lung cancer—will determine whether offering the policy will lead to a positive or negative outcome for the insurance company. How much should the analyst pay to learn whether or not the potential customer smokes?

Let us begin by explicating the decision problem that the analyst faces. They have to choose between two actions: Approve or Deny. The utility that they get from performing each of these actions depends on whether or not the potential customer will develop lung cancer. Table 3.1 specifies the utilities of each action in each relevant state of the world. The analyst cannot learn whether the customer will develop lung cancer, but can learn whether or not the customer smokes. Suppose further that if a person smokes, then there is a .1 probability that they will develop lung cancer. If they do not smoke, then there is a .01 probability that they will develop lung cancer. Assume that each of these probabilities are "causal" probabilities, in the

sense that the value of the causal variable is set via an intervention. Assume further that an intervention will be performed on the causal variable; the analyst is just unsure as to which intervention will be performed. Now suppose that the analyst learns that the potential customer does in fact smoke. Under these conditions, the expected utility of Approve is $.1(-100)+.9(10) = -1$ and the expected utility of Deny is 0. So the analyst would choose to Deny, and expect a utility of 0. Next, suppose that the analyst learns that the person does not smoke. Under this supposition, the expected utility of Approve would be $.01(-100) + .99(10) = 8.9$, and the expected utility of Deny would be 0. So the analyst would choose to Approve, and expect a utility of 8.9. Finally, let us suppose that the analyst has the prior belief that a person is equally likely to be intervened upon so as to smoke or not smoke (to be clear, this is a further assumption, rather than an implication of anything else stipulated so far). Under this supposition, the expected utility of learning which intervention has been performed on the potential customer's smoker status is $.5(0) + .5(8.9) = 4.45$.

However, to calculate the fair price of information about whether or not the potential customer smokes, one also has to consider what the analyst would have done had they not learned the potential customer's smoker status. Their prior beliefs about whether the potential customer smokes and their posterior beliefs about whether the potential customer develops lung cancer (given the customer's smoker status) jointly imply that the potential customer has prior probability .055 of developing lung cancer, and therefore a prior probability .945 of not developing lung cancer. Thus, if the analyst does not learn whether the customer smokes, then the expected utility of Approve is $.055(-100) + .945(10) = 3.95$ and the expected utility of Deny is 0. Since the analyst always acts to maximize expected utility, in the absence of information as to the potential customer's smoker status, they would choose Approve. If we subtract the expected utility of Approve when smoker-status is unknown from the expected utility of learning whether the potential customer smokes, we get $4.45 - 3.95 = .5$. This is the fair price of information regarding whether or not the potential customer smokes; the analyst should only pay up to half a unit of whatever currency the utilities are expressed in to learn whether or not the potential customer smokes.[3]

In some cases, refining the causal variable adds no pragmatic value to a causal model. Consider the case from the previous section, in which a smoker has a probability $.1$ of developing lung cancer regardless of the brand that they smoke, and a non-smoker has a $.01$ chance of developing lung cancer. The utility matrix for the decision to approve or deny a customer for life insurance is the same as in Table 3.1. If we suppose that the brand smoked by the potential customer makes no difference to their probability of developing lung cancer, then we know that learning this information will not change the action with maximum expected utility, as compared

[3] See the appendix for an abstract model of the fair price of causal information.

to when the customer smokes. Thus, the FPCI with respect to whether the customer smokes Marlboros, another brand, or does not smoke will be the same as the FPCI with respect to whether or not the customer smokes. As there is no loss of pragmatic value at the more coarse-grained level of description, choosing the fine-grained variable over the coarse-grained variable introduces a level of complexity to the model that is worthless, from a pragmatic perspective, and therefore the refinement ought to be avoided.[4]

We do not need to worry here about the prior probability distribution over either causal variable. If it is the case that an agent would pay more to learn the value of the fine-grained variable than they would to learn the value of its coarse-grained counterpart, then the prior probability distribution over causal variables determines to some degree *how much* more an agent should pay for information about the value of a more fine-grained variable as compared to its coarse-grained counterpart. However, changes in the prior distribution over the causal variable will never undo this inequality. Similarly, if information about the value of a fine-grained variable is worth no more to an agent than information about the value of a coarse-grained version of the same variable, this equality will be invariant across changes to the probability distribution over either variable.

In general, we want it to be the case that if refining the causal variable makes no difference to the probability distribution over the effect variable, then the coarse-grained explanation is to be preferred to the fine-grained explanation on pragmatic as well as explanatory grounds. This should hold regardless of the agent's pragmatic interests. The FPCI measure defined here satisfies this desideratum. Let $\pi(u(\cdot), A, E, C)$ denote the FPCI for a causal variable C, given a utility function $u(\cdot)$, action set A and effect variable E. The following proposition is true:

Proposition 9. *For any causal variables C and C', effect variable E, utility function $u(\cdot)$ and probability distribution $p(\cdot)$, where C' and C are both causes of E and C is a coarsening of C', if $p(e_i|do(c'_l)) = p(e_i|do(c_j))$ for all sets of values e_i, c_j, and c'_l such that $C' = c'_l$ implies $C = c_j$, then $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E, C')$.*

This means that in the class of "easy cases" in which refining the causal variable makes no difference to the probability distribution over the effect, it is also the case that the coarse-grained explanation provides equal pragmatic value with less complexity, and can therefore be chosen over a fine-grained alternative. Of course, these are the same easy cases in which a more proportional explanation is also more abstract, and also intuitively deeper. What we need to do now is consider those cases in which a more proportional or more abstract explanation seems less deep

[4] In what follows, I will discuss cases in which refining the causal variable does make a small difference to the probability distribution over the effect variable.

	No Recovery	Recovery
Approve	-100	10
Deny	0	0

Tab. 3.2: Utility Matrix For Life Insurance Decision

than a less proportional or less abstract rival, and determine whether an FPCI-based measure of the pragmatic value of an explanation is of any help.

First, we can consider a case where proportionality and depth come apart. Here, the probability of developing lung cancer from smoking Marlboros was .1001, and the probability of developing lung cancer from smoking other brands was .0999. If we again use the utility matrix in Table 3.1, one can check that learning whether the potential customer smokes Marlboros, another brand, or does not smoke will not make any difference to the analyst's decision, as compared to simply learning whether or not the customer smokes. Thus, they will pay no more to learn the value of this more fine-grained variable than they would to learn the value of its coarse-grained counterpart. By considering the pragmatic value of causal information in assessing explanatory depth, we can conclude that the coarse-grained explanation 'the patient developed lung cancer because they smoked' is deeper than the fine-grained explanation 'the patient developed lung cancer because they smoked Marlboros', even in cases where Marlboros are slightly more carcinogenic than other cigarettes.

Next, consider the case of treating a patient with antibiotics that may or may not contain penicillin. Recall that in this case, a less abstract explanation seems to be deeper. Again, we can consider the decision faced by a life insurance analyst, where the analyst's utilities in the state of the world where the patient does or does not recover from the bacterial infection are given in Table 3.2. The relevant expected utilities for the analyst are as follows.

$$eu(\text{Approve}|\text{Penicillin}) = .05(-100) + .95(10) = 4.5 \tag{3.1}$$

$$eu(\text{Approve}|\text{Non-Penicillin Antibiotic}) = .3(-100) + .7(10) = -23 \tag{3.2}$$

$$eu(\text{Approve}|\text{No Drugs}) = .8(-100) + .2(10) = -78 \tag{3.3}$$

Once again the expected utility of Deny is always 0. Let us assume that the probability that a person receives antibiotics with penicillin is $.4$, the probability that a person receives antibiotics with penicillin is $.4$, and the probability that a person receives no treatment is $.2$. This entails that the prior probability of No Recovery is $.32$ and the prior probability of Recovery is $.68$, so that the expected utility of Approve in the absence of information about the patient's treatment is $.32(-100) + .68(10) = -25.2$. Thus, the fair price of causal information regarding the patient's treatment is $(.4(4.5) + .4(0) + .2(0)) - 0 = 1.8$.

If we coarse-grain the causal variable and consider only whether or not the patient receives antibiotics, the model loses pragmatic value. One can check that if the analyst knew that the patient had received antibiotics, but not whether those antibiotics contained penicillin, then the analyst would choose Deny. Since the analyst would also choose Deny if they learned that the patient did not receive antibiotics, and would choose Deny in the absence of information, there is no value to learning whether or not the patient received antibiotics; the fair price of causal information is zero. Thus, the coarse-grained explanatory model of the patient's recovery, i.e. the model that makes no mention of penicillin, is less pragmatically valuable than the fine-grained model that does mention penicillin. This goes some way towards providing a rationalization of why greater abstraction does not seem to result in greater explanatory depth in this case; the more abstract explanation conceals information that rises above some threshold of pragmatic usefulness. Thus, my pragmatic approach to explanatory depth proves useful in this case as well.

I do not claim that FPCI constitutes a totalizing measure of explanatory depth; I find it unlikely that such a measure could exist. However, I believe that FPCI addresses an important issue in the literature on explanatory depth. Woodward (2010), Ylikoski and Kuorikoski (2010) and Weslake (2010) all accept that the depth of a causal explanation at a particular level of granularity, as compared to explanations of the same phenomenon at finer or coarser grains, is sensitive to our interests as agents. However, this is often described in a very general way, without much mathematical precision. My proposal makes the interest-relative dimension of explanatory depth precise, by mathematically linking it to the utilities that an agent assigns to various outcomes in a given decision problem. More precisely, my proposal quantifies the extent to which an explanatory causal model enables an agent to make the best decision possible, as defined by the agent's own utility function over outcomes, while minimizing cognitive burdens due to the complexity with which the explanans is described. This stands in contrast to purely epistemic measures of explanatory depth, which consider only the probability distribution over the possible values of an explanandum, given the possible values of the explanans.

3.6 Potential Objections

There are several potential objections to the view presented above. First, one could worry that it is strange to consider the value of information in cases where we already possess the information in question. In the smoking case, we may already know that the patient is a Marlboro smoker, and in the antibiotics case, we may already know that the patient has been prescribed antibiotics containing penicillin. So why does it matter what the insurance analyst would pay to learn this information in a counterfactual scenario where they did not already know it?

One way of replying to this worry is to say that our choice of granularity for an explanatory model guides future searches for explanatory information. For example, when we say of a given patient that smoking caused their lung cancer, this implies that when we search for an explanation of some other patient's lung cancer, we can begin by asking "did they smoke?" Clearly, the answer to this question will be valuable to us as both explanatory and pragmatic agents. Now suppose instead that we ask, "did they smoke Marlboros?" If the answer is "no", then we will need to ask another question, namely, "did they smoke any other brands?" Regardless of the answer to this question, it is clear that it would have been better just to ask, in the first place, "did they smoke?" By contrast, in the antibiotics case, it seems that the answer to the question "was the patient given penicillin?" will be explanatorily useful to us in understanding why the patient recovered, regardless of whether the answer is "yes" or "no." In this case, the introduction of a finer granularity of causal information seems to provide a meaningful direction of inquiry in future cases, rather than merely introducing unnecessary complexity to scientific practice.

Next, one could worry that in the examples that I have given above, the interventionist theory of causation is actually doing very little work. For the life insurance analyst in question, information about the causes of lung cancer will not necessarily be any more valuable than information about non-causal correlates of lung cancer. To illustrate, it might be the case that information about how yellow a customer's fingers and teeth are as valuable to our insurance analyst as information about the customer's smoker status, even though yellow teeth and fingers are not causes of lung cancer. Further, depending on the structure of the broader causal model in which a given cause-effect relation is embedded, data about non-causal correlates may be *more* useful to an agent than data about the causes of some variable of interest. Finally, given that the interventions used in philosophical analyses of causation are usually understood as counterfactual, a real-life insurance analyst never *could* learn that such an intervention had been performed. So it is worth asking whether it makes sense to use an information-pricing model to measure the pragmatic value of a causal explanation.

In response, I concede that the example of an insurance analyst deciding how much to pay for information about a potential customer is not a perfect analogy, for exactly the reasons given above. However, the analogy still does serve its purpose. My project in this chapter is to provide a strategy for determining the deepest level of causal explanation in a given case. After determining that competing causal explanations at varying levels of granularity are all causal explanations of the same phenomenon, I then use an information-pricing analysis to settle the question of the optimal depth. This second step is not, in itself, uniquely suited to causal as opposed to non-causal models. However, my analysis takes for granted that we are already talking about causal explanations. For example, on my account we would first use an interventionist criterion of causation to determine that smoking, and not yellow teeth and fingers, is a cause of lung cancer. Then, we would use an FPCI-based criterion to determine the optimal level of granularity for describing a person's smoker status when explaining why they have lung cancer. Variables denoting the color of a person's teeth or fingers do not make it to this second stage of analysis, since they are not causes of lung cancer in the first place.[5]

Finally, there is some force to the anti-pragmatic notion that the better explanation of some phenomenon should never depend on the context-specific utilities of the agent doing the explaining. To put it bluntly, it seems strange to think that life-insurance premiums should influence what counts as the best explanation of a patient's lung cancer. However, attempts up to this point to give some purely non-pragmatic account of the value of explanations have proven unsuccessful. As Franklin-Hall (2016) points out, the space of possible variables that we can use to represent a given causal explanans is enormous, presenting reasons for pessimism that a single epistemic criterion could rule out all but the best variables for a given explanation in the interventionist framework. The abstraction criterion proposed by Weslake (2010) solves some problems, but runs into others, both because it allows disjunctive variables—as Weatherson (2012) and Clarke (2017) have pointed out—and in probabilistic cases, as I have pointed out here. The same goes for proportionality, in part for reasons described above, and in part for reasons described by Shapiro and Sober (2012) that I have addressed in Chapter 1. Finally, Franklin-Hall (2016) argues that analyzing explanatory depth in terms of the stability of an explanation, i.e. its robustness in the face of changes to background conditions, also allows for putatively deep explanations that use disjunctive variables. The use of such variables seems out of keeping with scientific practice, and thus stability also fails as an epistemic criterion for explanatory depth.

[5]Having said this, an interesting avenue for future research would be an exploration of the overlap between information-pricing analyses of explanatory models and existing work on the expected utility of interventions and observations in inferring causal graphs (see Eberhardt 2007, p. 170).

In light of this failure to find an epistemic criterion for explanatory depth, I believe that we ought to take seriously the idea that the notion of superiority that we attach to some explanations is fundamentally an artifact of our perspective as pragmatically-motivated agents, rather than some observer-independent feature of nature. As such, the intrusion of pragmatics into judgments of explanatory depth may be inevitable, especially in threshold cases where different explanatory desiderata such as proportionality and abstractness pull in different directions. My proposal provides a precise way to take into account an agent's prudential interests in order to give an account of how these interests determine an optimal granularity for causal explanations. This pragmatic proposal for determining the optimal granularity for describing a causal relation holds regardless of whether we follow van Fraassen (1980) in taking explanation to be fundamentally pragmatic, or follow Salmon (1998) in taking explanatory relations to be an objective relation between events or objects.

There is some precedent for this kind of approach in the literature. Potochnik (2015, 2017) and Andersen (2017) follow Dennett (1991) in arguing that causal relata are high-level "patterns" formed of lower-level "pixels", where patterns are realized by pixels, but are robust over a certain amount of noise in their pixel-level realizations. To use Dennett's example, a bar code composed of individual pixels may cause a scanner to receive a certain message when the bar code is scanned. This causal relationship between the bar code and the scanner may be robust over many possible changes in the individual pixels comprising the bar code. The degree of noise that we will tolerate while still picking out the bar code as an important causal relatum is set in part via pragmatic concerns; as long as the bar code does its job, we can speak coherently of the bar code playing an explanatory causal role with respect to the transfer of information, and not worry about the particular arrangement of pixels. That this account of higher-level causation is sensitive to pragmatic concerns is explicated by Andersen when she writes that "the more efficiently we describe a pattern, the faster we can identify whether it occurs. If there are tasks for which speed is relevant, we might prioritize efficiency of description and accept reduced accuracy" (2017, p. 603). My proposal here makes a similar point; our pragmatic concerns clearly play some role in the kinds of events that we pick out as explanatorily salient, and the level of granularity with which we pick them out.

3.7 Conclusion

This chapter has considered the way in which proportionality and abstractness provide an account of explanatory depth in deterministic cases. I then show how this account unravels in the probabilistic context. I propose a novel measure of

the pragmatic value of an explanatory framework, and show how it addresses the problematic cases considered. I argue that this measure enables us to account for the intuition that some explanations are better than others in probabilistic contexts, albeit via the essential consideration of the pragmatic context of an explanation.

3.8 Appendix

3.8.1 General Definition of FPCI

Begin with set of acts $A = \{a_1, a_2, ..., a_r\}$, a value set $\{e_1, e_2, ..., e_n\}$ for the effect variable E, and a value set $\{c_1, c_2, ..., c_m\}$ for the causal variable C. Let $u(\cdot)$ be a utility function that takes as its arguments an action a_k and a value of the effect variable e_i. Let $p(\cdot)$ be a joint probability function over the values of E and C. We can define the joint probability matrix $\mathbf{P_{E,C}}$, utility matrix $\mathbf{U}$, and effect vector $\vec{P_E}$ as follows.

$$\mathbf{P_{E,C}} = \begin{bmatrix} p(e_1|do(c_1))p(do(c_1)) & p(e_1|do(c_2))p(do(c_2)) & ... & p(e_1|do(c_m))p(do(c_m)) \\ p(e_2|do(c_1))p(do(c_1)) & p(e_2|do(c_2))p(do(c_2)) & ... & p(e_2|do(c_m))p(do(c_m)) \\ \vdots & \vdots & \vdots & \vdots \\ p(e_n|do(c_1))p(do(c_1)) & p(e_n|do(c_2))p(do(c_2)) & ... & p(e_n|do(c_m))p(do(c_m)) \end{bmatrix} \tag{3.4}$$

$$\mathbf{U} = \begin{bmatrix} u(a_1, e_1) & u(a_1, e_2) & ... & u(a_1, e_n) \\ u(a_2, e_1) & u(a_2, e_2) & ... & u(a_2, e_n) \\ \vdots & \vdots & \vdots & \vdots \\ u(a_r, e_1) & u(a_r, e_2) & ... & u(a_r, e_n) \end{bmatrix} \tag{3.5}$$

$$\vec{P_E} = \begin{bmatrix} p(e_1) \\ p(e_2) \\ \vdots \\ p(e_n) \end{bmatrix} \tag{3.6}$$

Let $max(\cdot)$ be a function that takes a matrix as its argument, and returns a vector containing the maximum value in each column of that matrix, with vertical vectors treated as one-column matrices. Let $sum(\cdot)$ be a function that takes a vector as its argument and returns the sum of each element in the vector. We can define the fair price of causal information for a causal variable C, utility matrix $\mathbf{U}$, and effect variable E as follows.

$$\pi(u(\cdot), A, E, C) = sum(max(\mathbf{U} \cdot \mathbf{P_{E,C}})) - max(\mathbf{U} \cdot \vec{P_E}) \tag{3.7}$$

3.8.2 Proof of Proposition 9

Proof. Begin by writing the formula for the FPCI of causal variables C and C':

$$\pi(u(\cdot), A, E, C) = sum(max(\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}})) - max(\mathbf{U} \cdot \vec{P_E}) \tag{3.8}$$

$$\pi(u(\cdot), A, E, C') = sum(max(\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'})) - max(\mathbf{U} \cdot \vec{P_E}) \tag{3.9}$$

Next, we can show that in this case, $sum(max(\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'})) = sum(max(\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}}))$. To do so, we begin by expanding the matrices $\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}}$ and $\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'}$.

$$\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}} = \begin{bmatrix} p(do(c_1)) \sum_{i=1}^{n} p(e_i|do(c_1))u(a_1, e_i) & \dots & p(do(c_m)) \sum_{i=1}^{n} p(e_i|do(c_m))u(a_1, e_i) \\ \vdots & \vdots & \vdots \\ p(do(c_1)) \sum_{i=1}^{n} p(e_i|do(c_1))u(a_r, e_i) & \dots & p(do(c_m)) \sum_{i=1}^{n} p(e_i|do(c_m))u(a_r, e_i) \end{bmatrix} \tag{3.10}$$

$$\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'} = \begin{bmatrix} p(do(c'_1)) \sum_{i=1}^{n} p(e_i|do(c'_1))u(a_1, e_i) & \dots & p(do(c'_q)) \sum_{i=1}^{n} p(e_i|do(c'_q))u(a_1, e_i) \\ \vdots & \vdots & \vdots \\ p(do(c'_1)) \sum_{i=1}^{n} p(e_i|do(c'_1))u(a_r, e_i) & \dots & p(do(c'_q)) \sum_{i=1}^{n} p(e_i|do(c'_q))u(a_r, e_i) \end{bmatrix} \tag{3.11}$$

Next, consider any value c_j such that $C = c_j$ if and only if C' takes a value in some set $\{c'_l, c'_{l+1}, ..., c'_{l+z}\}$. We can define submatrices $\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}}[j]$ and $\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'}[l : l + z]$ as follows.

$$\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}}[j] = \begin{bmatrix} p(do(c_j)) \sum_{i=1}^{n} p(e_i|do(c_j))u(a_1, e_i) \\ \vdots \\ p(do(c_j)) \sum_{i=1}^{n} p(e_i|do(c_j))u(a_r, e_i) \end{bmatrix} \tag{3.12}$$

$$\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'}[l : l + z] = \begin{bmatrix} p(do(c'_l)) \sum_{i=1}^{n} p(e_i|do(c'_l))u(a_1, e_i) & \dots & p(do(c'_{l+z})) \sum_{i=1}^{n} p(e_i|do(c'_{l+z}))u(a_1, e_i) \\ \vdots & \vdots & \vdots \\ p(do(c'_l)) \sum_{i=1}^{n} p(e_i|do(c'_l))u(a_r, e_i) & \dots & p(do(c'_{l+z})) \sum_{i=1}^{n} p(e_i|do(c'_{l+z}))u(a_r, e_i) \end{bmatrix} \tag{3.13}$$

The sum of every element in a given row of $\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'}[l : l + z]$ equals the element in the corresponding row of $\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}}[j]$. If $p(e_i|do(c'_l)) = p(e_i|do(c_j))$ for all sets of values e_i, c_j and c'_l such that $C' = c'_l$ implies $C = c_j$, then $\sum_{i=1}^{n} p(e_i|do(c'_{l+v}))u(a_k, e_i) =$

$\sum_{i=1}^{n} p(e_i|do(c_j))u(a_k, e_i)$ for any a_k. This implies that the submatrix $\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'}[l : l + z]$ is row-dominated. It follows from this that $sum(max(\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'}[l : l + z])) = sum(max((\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}}[j]))$ for any c_j and any set $\{c'_l, c'_{l+1}, ..., c'_{l+z}\}$ such that $\{c'_l, c'_{l+1}, ..., c'_{l+z}\}$ is the union of all values of C' that imply that $C = c_j$. It follows from this that $sum(max(\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}'})) = sum(max(\mathbf{U} \cdot \mathbf{P}_{\mathbf{E},\mathbf{C}}))$. Since $max(\mathbf{U} \cdot \vec{P_E})$ is unchanged by refining the causal variable, this implies that $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E, C')$. □

4 Pragmatic Causal Feature Learning

4.1 Introduction

The previous chapter proposed a methodology for coarsening a causal variable in a way that only preserves pragmatically salient information. Chalpuka et al. (2015, 2016, 2016, 2017), using techniques developed by Shalizi and Crutchfield (2001), propose an alternative coarsening procedure, which they call "causal feature learning". To my knowledge, their procedure is the primary contribution to the machine learning literature on the issue of coarsening variables in Bayesian networks. Chalupka et al.'s approach to coarsening differs from my proposal in an important way. Specifically, it focuses on the epistemic, rather than pragmatic, value of causal information, in a sense that I will make clear in this chapter.

Chalupka et al. emphasize two good-making features of their procedure. First, their procedure returns a coarsening of both a causal variable and its effect variable, so as to describe an entire cause-effect pair in a coarse-grained way that preserves all and only the epistemically valuable information shared between the variables. By contrast, in the previous chapter I only considered possible coarsenings of the causal variable. Second, Chalupka et al. show that their method can be applied to observational data, rather than experimental data, to produce a coarsening that preserves all and only the epistemically valuable causal information shared between the two variables, with this procedure only failing in a Lebesgue measure-zero subset of possible cases. They call this result the "causal coarsening theorem". The details and significance of this theorem are explained more thoroughly in what follows.

It would reflect poorly on the procedure proposed in the previous chapter if, unlike Chalupka et al.'s approach, it could not be extended to provide guidance on how to coarsen effect variables. It would also be a bad result for my procedure if it could not be applied to observational data, with failure in only a Lebesgue measure-zero subset of cases. Thankfully, I am able to avoid these bad results. In this chapter, I show that my pragmatic coarsening procedure retains both of the positive features of Chalupka et al.'s approach outlined above.

The plan for this chapter is as follows. In Section 4.2, I provide an exegesis of Chalupka et al.'s framework for causal feature learning. In Section 4.3, I state their causal coarsening theorem and discuss its significance. In Section 4.4, I recap my pragmatic approach from the previous chapter, framing it in a way that allows for direct comparison with Chalpuka et al.'s approach to causal feature learning. In Section 4.5, I consider and rebut potential arguments against my proposal. In Section 4.6, I offer concluding remarks.

4.2 Chalupka et al.'s Framework

Chalupka et al.'s proposal can be summarized as follows. Let $\mathcal{N}' = \langle \mathcal{V}' \cup \mathcal{I}^{\mathcal{V}'}, \mathcal{E}', p(\cdot) \rangle$ be a Bayes net that satisfies CMC and Minimality, and which is an "extended" Bayes net that includes intervention variables, as discussed in Section 1.2.3. As in Chapter 1, a superscript $\prime$ denotes a more fine-grained variable or a set or graph containing more fine-grained variables. Let $C' \in \mathcal{V}'$ be an ancestor of another variable $E' \in \mathcal{V}'$. Recall also that, in light of Huang and Valtorta's (2006) result, we can use the graphical structure of $\mathcal{N}'$ to calculate the interventional conditional probability distribution over every variable in the network, given an intervention on any other variable or group of variables in the network. These preliminaries allow us to define the first crucial equivalence relation in Chalupka et al.'s framework. The *epistemic causal equivalence* relation is defined over the value set of the variable C', in relation to the variable E', and is stated as follows:

> **Epistemic Causal Equivalence**: For any two values c'_l and c'_t, $c'_l \sim_{ec} c'_t$ if and only if for all e'_s, $p(e'_s|do(c'_l)) = p(e'_s|do(c'_t))$.

In other words, Chalupka et al. consider two fine-grained values of a causal variable C' to be equivalent if and only if an intervention setting C' to either value results in the same probability distribution over the effect variable E'. Note that while Chalupka et al. call their equivalence relation simply 'causal equivalence', I attach the label 'epistemic' to the formal concepts used in their framework, so as to distinguish them from the pragmatic concepts discussed in my framework.

This equivalence relation allows Chalupka et al. to define a variable C^{ec} that is an *epistemic causal coarsening* of the fine-grained causal variable C':

> **Epistemic Causal Coarsening:** C^{ec} is an epistemic causal coarsening of C' if and only if the value set of C^{ec} is a quotient set of the value set of C' according to the equivalence relation $\sim_{ec}$.

Recall from Chapter 1 that a set A is a quotient set of another set B just in case each element of A is an equivalence class of the elements in B according to some equivalence relation defined over B, and A is a partition of B. Thus, each value of C^{ec} is a subset of the set of values of C', each element of which stands in the equivalence relation $\sim_{ec}$ to every other element of the subset.

Next, Chalupka et al. define the following equivalence relation over the value set of the fine-grained effect variable E', in relation to the variable C':

> **Epistemic Effect Equivalence**: For any two values e'_s and e'_u, $e'_s \sim_{ee} e'_u$ if and only if for all c'_l, $p(e'_s|do(c'_l)) = p(e'_u|do(c'_l))$.

In other words, two values of an effect variable are equivalent if and only if any intervention on the causal variable results in those two values of the effect variable having the same probability. This in turn yields the following definition of an epistemic effect coarsening:

> **Epistemic Effect Coarsening:** E^{ee} is an epistemic effect coarsening of E' if and only if the value set of E^{ee} is a quotient set of the value set of E' according to the equivalence relation $\sim_{ee}$.

In Section 4.5, I will discuss what I take to be some negative features of this way of coarsening the effect variable, but for now we have in place the formal concepts needed to summarize Chalupka et al.'s procedure for causal feature learning.

For Chalupka et al., causal feature learning proceeds as follows. First, identify a cause-effect pair (C', E') within a Bayes net $\mathcal{N}' = \langle \mathcal{V}', \mathcal{E}', p(\cdot) \rangle$. Then, for each value c'_l of C', exploit the joint probability distribution $p(\cdot)$ and the graphical structure of $\mathcal{N}'$ to identify all other values c'_t such that $c'_l \sim_{ec} c'_t$. Then, for each value e'_s of E', exploit the joint probability distribution $p(\cdot)$ and the graphical structure of $\mathcal{N}'$ to identify all other values e'_u such that $e'_s \sim_{ee} e'_u$. Finally, use these equivalence relations to obtain the coarsenings C^{ec} and E^{ee}. For a given Bayes net $\mathcal{N}'$ and cause-effect pair (C', E'), the pair of coarsenings (C^{ec}, E^{ee}) is unique (Chalupka et al. 2017, 143).

In the previous chapter, I discussed Pocheville et al.'s (2017) criterion for determining when a probabilistic causal relationship between a causal variable and its effect is maximally *proportional*. According to this criterion, a cause-effect relationship is maximally proportional when each possible intervention on the causal variable results in a different probability distribution over the effect variable. It should be clear that Chalupka et al.'s method for causal feature learning results in a coarse-grained cause and effect variable that together satisfy this desideratum. In the

previous chapter, I argued that proportionality, so defined, is not necessarily an ideal to aim for when coarsening variables in a probabilistic causal model. Thus, it is not surprising that I put forward an alternative to Chalupka et al.'s coarsening strategy in this chapter as well, and that pragmatic considerations once again play a central role in my alternative proposal.

4.3 The Causal Coarsening Theorem

Chalupka et al.'s principle formal achievement is to prove that, except in a Lebesgue measure-zero subset of cases, the epistemic causal coarsening and epistemic effect coarsening of a given variable can be learned from observational rather than experimental data. To state their result in a perspicuous way, we first define additional equivalence relations and coarsenings over the variables in the cause-effect pair (C', E'). Let us begin with the *observational causal equivalence* relation:

> **Observational Causal Equivalence**: For any two values c'_l and c'_t, $c'_l \sim_{oc} c'_t$ if and only if for all e'_s, $p(e'_s|c'_l) = p(e'_s|c'_t)$.

Note that the sole difference between observational causal equivalence and epistemic causal equivalence is that the former is defined using observational conditional probabilities (i.e. conditional probabilities such that the conditioning event is *not* set via an intervention), whereas epistemic causal equivalence is defined using causal conditional probabilities (i.e. the conditioning event *is* set via an intervention). This leads directly to the following definition of an observational causal coarsening:

> **Observational Causal Coarsening:** C^{oc} is an observational causal coarsening of C' if and only if the value set of C^{oc} is a quotient set of the value set of C' according to the equivalence relation $\sim_{oc}$.

Similar definitions of an observational equivalence relation and accompanying coarsening can be given for the effect variable:

> **Observational Effect Equivalence**: For any two values e'_s and e'_u, $e'_s \sim_{oe} e'_u$ if and only if for all c'_l, $p(e'_s|c'_l) = p(e'_u|c'_l)$.

> **Observational Effect Coarsening:** E^{oe} is an observational causal coarsening of E' if and only if the value set of E^{oe} is a quotient set of the value set of E' according to the equivalence relation $\sim_{oe}$.

These observational coarsenings of the fine-grained cause and effect variable are ways of coarsening these variables using only observational conditional probabilities, without any conditioning on hypothetical interventions.

The causal conditional probability distribution over some variable E given any value of C can differ from the observational conditional probability distribution over some variable E given any value of C. This divergence can be due to some confounding variable Z that is a cause of both C and E. For instance, suppose that cardiovascular exercise causes weight loss, and that drinking caffeine causes both increased cardiovascular exercise and increased weight loss. The observation that someone engages in cardiovascular exercise will be strongly correlated with the observation that they lose weight, since the observation of cardiovascular exercise is informative as to weight loss both because of the direct causal relationship between exercise and weight loss and because an observation of exercise is evidence of caffeine consumption. By contrast, in this scenario, weight loss may be less strongly correlated with interventions increasing the amount of exercise that a person does, since the exercise amounts set via intervention are not causally related to caffeine consumption. In another typical confounding scenario, Z may be an effect of C and a cause of E. For instance, suppose that exercise causes weight loss, but also additional eating, which leads to weight gain. The result is that exercisers do not seem to lose much weight, if any, and may even gain weight. Thus, the amount that a person eats confounds the causal relationship between exercise and weight. An observation that someone is exercising more may not change our beliefs about their potential weight loss very much, but learning that someone has been forced to exercise, holding fixed other factors, may change our beliefs about their potential weight loss considerably.

Chalupka et al. argue that when we coarsen fine-grained cause and effect variables, we can almost always ignore the distorting influence of potential confounders. At the crux of their argument is the following proposition:

Proposition 10 (Causal Coarsening Theorem). *Let C', E', and Z' be variables in a graph $\mathcal{G}'$ such that C' is an ancestor of E' and Z' is a possible confounder of the causal relationship between C' and E'. Consider the set of possible joint probability distributions over C', E', and Z'. Let C^{ec} and C^{oc} be the epistemic causal coarsening and observational causal coarsening of C', respectively, in any such probability distribution. Let E^{ec} and E^{oc} be the epistemic effect coarsening and observational effect coarsening of E', respectively, in any such probability distribution. The set of joint probability distributions over C', E', and Z' such that the value set of C^{ec} is not a quotient set of the value set of C^{oc} and the value set of E^{ec} is not a quotient set of the value set of E^{oc} is Lebesgue measure zero within the set of all possible joint distributions over C', E', and Z'.*

To give an overview of the proof of Chalupka et al.'s result, suppose that the variable Z' has cardinality k, C' has cardinality m, and E' has cardinality n. Each possible joint probability distribution over these three variables can be represented as a point in the space $\mathbb{R}^{(n-1)\times k^2(k-1)\times m(m-1)}$. The set of all such points forms a simplex in that space. Chalupka et al. define polynomial constraints on all points in the simplex such that, if these constraints are satisfied, then C^{ec} is not a coarsening of C^{oc}, and E^{ec} is not a coarsening of E^{oc}. They then use a result from Okamoto (1973) to show that because these constraints are non-trivial, they only hold in a region of the simplex that has Lebesgue measure zero within the simplex.

This result has a putatively important practical consequence. Suppose that we want to find the epistemic causal coarsening of some fine-grained variable C' that is a cause of some other fine-grained variable E'. If we are guided solely by the coarsening strategies outlined above, then doing so would require a separate intervention for each value of C', since each causal conditional probability would be needed to determine which values of C' and E' are epistemically equivalent. However, equipped with Proposition 10, we can instead use a potentially more efficient procedure. Using just observational data, we can coarsen the micro-variable C' into its observational causal coarsening C^{oc}. We can then intervene to set C^{oc} to each of its values, and thereby determine which of its values are epistemically causally equivalent, thereby generating the epistemic causal coarsening C^{ec}. This procedure is potentially more efficient since C^{oc} never has more values than C', and may have many less values that C'. The procedure is justified just in case the joint probability distribution over C', E', and confounder Z' is such that C^{ec} is a coarsening of C^{oc}, a condition that provably only fails for a measure-zero set of probability distributions. We can give the same account of the usefulness of the causal coarsening theorem when the variable being coarsened is the effect variable E'.

Further, note that in Proposition 10, the full graphical structure of the Bayes net in which the variables C', E', and Z' are embedded is not fully specified. We only know that C' is an ancestor of E'. Thus, Chalupka et al. show that they are able to infer the causal coarsening of C' and E' without knowing the full causal structure of the system in question. This speaks to the significance of their result, since if we did know the graphical structure of the Bayes net in which C', E', and Z' are embedded, then the completeness results of Huang and Valtorta (2006) show that we could obtain causal conditional probabilities for all values of C' and E', thereby obtaining the epistemic causal coarsenings from observational data without error.

Just how positive a result this is for causal feature learning depends on how one views the practice of discounting the possibility of Lebesgue measure-zero events when making inductive inferences. If one believes it is a good principle of induction that measure-zero events can be assumed not to occur when making inductive

inferences, then Proposition 10 demonstrates a good-making feature of Chalupka et al.'s procedure. Alternatively, one might be skeptical of the practice of ignoring the possibility of measure zero events when making inductive inferences. For instance, one might follow Hájek (2003) in holding that measure-zero events are epistemically possible, and therefore cannot be discounted when making inductive inferences. If one adheres to this view, then the result stated in Proposition 10 is decidedly less consequential.[1]

In what follows, I remain neutral as to the epistemological significance of the Causal Coarsening Theorem. I argue only for the conditional claim that if Chalupka et al.'s framework for causal feature learning has the good-making features that its proponents take it to have—in virtue of the truth of the causal coarsening theorem—then my framework inherits these same good-making features. Specifically, I will show in Section 4.4.5 that my pragmatic framework for causal feature learning can also be applied when we only have observational probabilities and the only causal information that we know is that C' is an ancestor of E', with failure occurring only in a Lebesgue measure-zero subset of cases.

4.4 The Pragmatic Approach to Coarsening

4.4.1 The Primary Goal of the Framework

I begin by recapping the example from the previous chapter, with some minor amendments, in order to motivate my approach to pragmatic causal feature learning. Suppose that a life insurance analyst can learn some information about a potential client, and thereafter decide whether or not to sell the customer a policy. Specifically, the analyst can run a query that tells them whether the potential customer is a Marlboro smoker, a smoker of another brand, or a non-smoker. As it turns out, if the potential customer turns out to be a Marlboro smoker or a smoker of another brand, then the analyst will not sell them a policy, due to the justifiable fear that the potential customer will die prematurely.[2] If the customer is a non-smoker, then the analyst

[1] Tangential to this discussion is a debate over similar results in Meek (1995) and Spirtes et al. (2000) arguing that potential violations of the Faithfulness condition in a Bayesian network can be discounted in causal search, since the set of joint probability distributions over a network that result in violations of Faithfulness is Lebesgue measure zero. Indeed, Chalupka et al.'s proof of the causal coarsening theorem uses the same techniques as Meek's proof. As Steel (2006) points out, these results will not be convincing to those who hold that violations of Faithfulness are not only possible but common, such as Cartwright (1999) and Hoover (2001). Further, Lin and Zhang (2018) offer independent arguments against Meek-style proofs of Faithfulness, which can also be read as critiques of the epistemic significance of the causal coarsening theorem.

[2] Note that I have switched here from using the presence of lung cancer as the principle outcome of concern for the life-insurance agent to using age at death as the principle outcome of concern for the life-insurance agent.

will sell them a policy. In such a case, information about the brand of cigarette that the potential customer smokes is not pragmatically relevant to the analyst; it will not affect the decision that they make. Thus, the analyst can change the causal model that is informing their decision-making in the following way: prior to making their decision, instead of querying a variable with the set of values {Marlboro smoker, smoker of other brands, non-smoker}, they can simply query the binary variable {smoker, non-smoker}. Coarsening the three-valued variable into the binary variable makes no difference to the way in which the analyst will act, once they learn the value of either variable. As such, there is no loss of valuable information, relative to a particular agent (the life-insurance analyst) facing a particular problem (whether to sell the potential customer life insurance). To put this in the language of the previous chapter, coarsening the causal variable in this case does not decrease the *fair price of information* about the value of the causal variable.

In the previous chapter, I argued that considering the fair price of information about the value of a causal variable in a particular decision context allows us to explain some intuitive judgments about the optimal level of granularity for a causal explanation. Specifically, I held that we can choose a more coarse-grained causal variable over a more fine-grained one when the fair price of information about the coarse-grained causal variable is equal to the fair price of information about the fine-grained causal variable. Here, I want to define a procedure for identifying the coarsest possible version of *both* the causal variable *and* the effect variable such that pragmatic value of information about the causal variable is optimized when moving from fine-grained to coarse-grained variables. This discussion differs from the upshot of the previous chapter in two ways. First, it defines pragmatic coarsening in terms of equivalence relations between values of variables, following Chalupka et al.'s formal approach. Second, it considers how to coarsen not only the causal variable, but also the effect variable, so as to optimize the pragmatically valuable information that is shared between a cause and its effect.

4.4.2 Coarsening the Causal Variable

This subsection largely translates my proposal in the previous chapter into the formal language of Chalupka et al.'s framework. I begin with some formal preliminaries. Let the variable C' be an ancestor of E' in a Bayes net $\mathcal{N}'$. Let A be a set of actions that the agent can choose to perform. Finally, let $u(\cdot)$ be a real-valued utility function over the cross-product of A and the value set of E'. This utility function, along with the graphical structure and joint probability distribution over C' and E' specified by $\mathcal{N}'$, allows us to calculate the expected utility of any action in A. Let $a_k \in A$ be any action, and let c'_l be some value of C'. Let w be the number of values in E'.

The expected utility of a_k, given an intervention setting C' to the value c'_l, is the following:

$$eu(a_k|do(c'_l)) = \sum_{s=1}^{w} u(a_k, e'_s)p(e'_s|do(c'_l)) \tag{4.1}$$

We define a pragmatic equivalence relation between values of the fine-grained causal variable C', in relation to the variable E', as follows:

> **Pragmatic Causal Equivalence**: For any two values c'_l and c'_t, $c'_l \sim_{pc} c'_t$ if and only if $\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c'_l)) = \underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c'_t))$.

In other words, two values c'_l and c'_t are pragmatic causal equivalents if and only if there is an action that has maximal expected utility when C' is set to either value via an intervention. For example, in the case above, because the same action (declining to sell life insurance) has maximum expected utility whether the potential customer is a Marlboro smoker or a smoker of other brands, the values of the causal variable 'Marlboro smoker' and 'smoker of other brands' are pragmatic causal equivalents.

In keeping with Chalupka et al.'s approach, we use this pragmatic causal equivalence relation to define a pragmatic causal coarsening of the fine-grained variable C':

> **Pragmatic Causal Coarsening:** C^{pc} is a pragmatic causal coarsening of C' if and only if the value set of C^{pc} is a quotient set of the value set of C' according to the equivalence relation $\sim_{pc}$.

In the example given above, the variable with the value set {smoker, non-smoker} is the pragmatic causal coarsening of the fine-grained variable with the value set {Marlboro smoker, smoker of other brands, non-smoker}.

There is a close connection between this definition of a pragmatic causal coarsening and the fair price of causal information. Recall from the previous chapter that $\pi(u(\cdot), A, E', C')$ denotes the price up to which an agent with utility function $u(\cdot)$ over the cross-product of a set of actions A and the value set of a variable E' would pay to learn which interventions had been performed on a variable C'. The following propositions are true:

Proposition 11. *If C^{pc} is the pragmatic causal coarsening of C', then it follows that $\pi(u(\cdot), A, E', C^{pc}) = \pi(u(\cdot), A, E', C')$.*

Proposition 12. *If C^{pc} is the pragmatic causal coarsening of a fine-grained causal variable C' and C is a variable such that the value set of C is a quotient set of C', then it follows that: 1) $\pi(u(\cdot), A, E', C) \leq \pi(u(\cdot), A, E', C^{pc})$, and 2) if $\pi(u(\cdot), A, E', C) = \pi(u(\cdot), A, E', C^{pc})$, then the value set of C^{pc} is a quotient set of the value set of C.*

	Death < 50	Death 50-70	Death > 70
Approve	-100	-70	10
Deny	0	0	0

Tab. 4.1: Utility Matrix For Life Insurance Decision

	Death < 50	Death 50-70	Death > 70
Marlboro Smoker	.003	.3	.697
Other Smoker	.002	.2	.798
Non-Smoker	.001	.05	.949

Tab. 4.2: Causal Conditional Probability Table

Proposition 11 tells us that an agent would pay just as much to learn the value of C^{pc} as they would pay to learn the value of C', all other things being equal. Proposition 12 tells us that replacing a given fine-grained variable C' with its pragmatic causal coarsening C^{pc} allows us to maximize the coarseness with which the causal variable is defined without reducing the fair price of causal information, provided that other factors are held fixed.

4.4.3 Coarsening the Effect Variable

In a departure from the previous chapter, we will now consider how to coarsen the effect variable so as to maximize the price of information about some causal variable, for an agent with a prudential interest in the value of the effect variable being coarsened. Here it is important to note that, under some reasonable assumptions, coarsening the effect variable can sometimes *increase* the fair price of information about the value of the causal variable. To illustrate, consider the following example. Table 4.1 shows the utility that an agent receives from approving or denying a potential customer for life insurance depending on whether or not the potential customer dies before 50, dies between 50 and 70, or dies after 70. Table 4.2 shows the conditional probability of each mortality category, given an intervention setting the customer's smoker status to each of three possible values. Under these conditions, and assuming a uniform prior probability distribution over possible smoker statuses, the fair price of causal information about the customer's smoker status is $1.96\bar{3}$ (see appendix for calculation).

However, suppose that we coarsen the effect variable so that its set of possible values is $\{\text{Death} \leq 70, \text{Death} > 70\}$. Suppose further that we wish to calculate the fair price of causal information with respect to the customer's smoker status, when the agent's utilities are now defined over the more coarse-grained effect

	Death ≤ 70	Death > 70
Approve	-70.32	10
Deny	0	0

Tab. 4.3: Utility Matrix For Life Insurance Decision

variable representing possible ages at death. In order to do this we have to answer two questions. First, how do we assign conditional probabilities to coarse-grained values of effect variables? Second, how do we assign utilities to each action when coarse-grained values of an effect variable obtain?

The first question is straightforwardly answered by the standard probability axioms. Let e_i be a value of a coarse-grained variable E that obtains if and only if the fine-grained variable E' takes a value in the set $\{e'_1, e'_2, ..., e'_\alpha\}$. For any value x_0 of any variable X, the following holds:

$$p(e_i|do(x_0)) = \sum_{s=1}^{\alpha} p(e'_s|do(x_0)) \tag{4.2}$$

This is just a special case of the countable additivity of probabilities of disjoint sets of events, i.e. Kolmogorov's third axiom.

As for the second question, there is no obvious reason to favor any one method of aggregating utilities when coarsening the space of events over which a utility function is defined. However, in what follows I will stipulate the following rule for such aggregation. Again, let e_i be a value of a coarse-grained variable E that obtains if and only if the fine-grained variable E' takes a value in the set $\{e'_1, e'_2, ..., e'_\alpha\}$. For any action a_k, the utility of performing a_k when $E = e_i$ is given by the following equation:

$$u(a_k, e_i) = \sum_{s=1}^{\alpha} \frac{p(e'_s)}{p(e_i)} u(a_k, e'_s) \tag{4.3}$$

In other words, the utility of performing an action a_k when a coarse-grained variable E takes some value e_i is equal to the weighted average of the utility of performing a_k when a fine grained variable E' takes each value in the set $\{e'_1, e'_2, ..., e'_\alpha\}$. The weights used in taking this average are calculated using the marginal probability distribution over the fine-grained variable E'.

Equipped with these aggregation rules, we can now coarsen the effect variable in the columns of Tables 4.1 and 4.2 to produce the utility matrix in Table 4.3 and the conditional probability distribution in Table 4.4 (see appendix for calculations). It turns out that if we use these coarse-grained tables to calculate the fair price of causal

	Death ≤ 70	Death > 70
Marlboro Smoker	.303	.697
Other Smoker	.202	.798
Non-Smoker	.051	.949

Tab. 4.4: Causal Conditional Probability Table

information about a customer's smoker status, we get the result that the agent should pay up to approximately 1.967 for this information (see appendix for calculation). Note that this price is strictly greater than the fair price of causal information about the customer's smoker status when the more fine-grained effect variable is used to represent possible ages at death. To see how this result is generated, consider the expected utility of approving a non-smoker for life insurance when the effect variable is fine-grained; it is $.001(-100) + .05(-70) + .949(10) = 5.89$. By contrast, the expected utility of approving a non-smoker for life insurance when the effect variable is coarsened is $.051(-70.32) + .949(10) \approx 5.90$. This slight increase in expected utility under the coarsening of the effect variable accounts for the slight increase in the fair price of information about the value of a causal variable.

Let us give a more general characterization of the increase in the fair price of causal information that occurs as a result of coarsening the effect variable in this case. Specifically, we can show that the following proposition is true:

Proposition 13. *Let $\{e'_1, e'_2, ..., e'_\alpha\}$ be a set of values of a fine-grained effect variable E'. Let e_i be a value of a coarse-grained effect variable E such that the value set of E is a quotient set of the value set of E' and $E = e_i$ if and only if E' takes a value in the set $\{e'_1, e'_2, ..., e'_\alpha\}$. For any utility function $u(\cdot)$, set of actions A, and causal variable C' with q values, $\pi(u(\cdot), A, E, C') \geq \pi(u(\cdot), A, E', C')$ if for all such e_i,*

$$\sum_{l=1}^{q} p(do(c'_l))p(e_i|do(c'_l))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e_i)$$

$$\geq \sum_{l=1}^{q} p(do(c'_l)) \sum_{l=1}^{\alpha} p(e'_s|do(c'_l))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e'_s).$$

For the sake of comprehension, note that $u(\operatorname{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e_i)$ is the utility that the agent receives when $E = e_i$ and the agent performs the action that maximizes expected utility after they learn that C' has been set to c'_l via an intervention.

Proposition 13 puts us in a position to define an equivalence relation over the value set of the effect variable, in relation to the value set of some causal variable, such

that coarsening according to this equivalence relation maximizes the fair price of causal information. This equivalence relation is defined as follows:

> **Pragmatic Effect Equivalence**: For any two values e'_s and e'_u, $e'_s \sim_{pe} e'_u$ if and only if, for a causal variable C' with q values:
>
> $$\sum_{l=1}^{q} p(do(c'_l))p(e'_s \vee e'_u|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e'_s \vee e'_u)$$
>
> $$\geq \sum_{l=1}^{q} p(do(c'_l))\Big(p(e'_u|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e'_s)$$
>
> $$+ p(e'_s|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e'_u)\Big).$$

Note that this inequality is just an instance of the inequality in Proposition 13, with the coarse-grained value e_i replaced by the disjunction $e'_s \vee e'_u$. This equivalence relation leads directly to the following definition of a pragmatic effect coarsening:

> **Pragmatic Effect Coarsening**: E^{pe} is a pragmatic effect coarsening of E' if and only if the value set of E^{pe} is a quotient set of the value set of E' according to the equivalence relation $\sim_{pe}$.

These definitions allow us to state the following proposition:

Proposition 14. *If E^{pe} is the pragmatic effect coarsening of a fine-grained effect variable E' and E is a variable such that the value set of E is a quotient set of the value set of E', then: 1) $\pi(u(\cdot), A, E, C') \leq \pi(u(\cdot), A, E^{pe}, C')$, and 2) if $\pi(u(\cdot), A, E, C') = \pi(u(\cdot), A, E^{pe}, C')$, then the value set of E^{pe} is a quotient set of the value set of E.*

That is, for any fine-grained effect variable E', the pragmatic effect coarsening E^{pe} is the coarsest coarsening of E' that maximizes the fair price of information about a causal variable C', for an agent with a utility function defined over E'.

4.4.4 The Pragmatically Optimal Coarsening

The goal of this section is to define a procedure for identifying, for a given scenario, the coarsest possible version of both the causal variable and the effect variable such that all pragmatically valuable information is preserved when moving from fine-grained to coarse-grained variables. I am now in a position to show that replacing a fine-grained causal variable C' with its pragmatic coarsening C^{pc} and replacing

a fine-grained effect variable E' with its pragmatic coarsening E^{pe} is just such a procedure. Specifically, the following proposition is true:

Proposition 15. *If C^{pc} is the pragmatic causal coarsening of C', E^{pe} is the pragmatic effect coarsening of E', the value set of C is a quotient set of the value set of C', and the value set of E is a quotient set of the value set of E', then: 1) $\pi(u(\cdot), A, E, C) \leq \pi(u(\cdot), A, E^{pe}, C^{pc})$, and 2) if $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E^{pe}, C^{pc})$, then 2a) the value set of C^{pc} is a quotient set of the value set of C and 2b) the value set of E^{pe} is a quotient set of the value set of E.*

This proposition combines the results of Propositions 11, 12, and 14 into a single result showing that C^{pc} and E^{pe} are the coarsest possible coarsenings of C' and E' that maximize the fair price of causal information about the causal variable, for an agent with a pragmatic interest in the effect variable.

In light of this result, pragmatic causal feature learning can be defined as follows. First, identify a cause-effect pair (C', E') within a Bayes net $\mathcal{N}' = \langle \mathcal{V}', \mathcal{E}', p(\cdot) \rangle$. Then, for each value of C' and E', exploit the joint probability distribution $p(\cdot)$ and the graphical structure of $\mathcal{N}'$ to identify the other values to which it stands in the relation $\sim_{pc}$ or $\sim_{pe}$. Finally, use these equivalence relations to obtain the coarsenings C^{pc} and E^{pe}. Additionally, the following proposition is true:

Proposition 16. *For any cause-effect pair (C', E'), action set A, utility function $u(\cdot)$ and joint probability distribution $p(\cdot)$, there is a unique pragmatic causal coarsening C^{pc} of C' with respect to E', and a unique pragmatic effect coarsening E^{pe} of E' with respect to C^{prime}, such that C^{pc} is the unique pragmatic causal coarsening of itself with respect to E^{pe}, and E^{pe} is the unique pragmatic effect coarsening of itself with respect to C^{pc}.*

Thus, pragmatic causal feature learning produces a unique result, which optimizes the fair price of information about the causal variable, for an agent with a prudential interest in the value of the effect variable.

The process of causal feature learning can be illustrated by our standard example of a life insurance agent querying a customer's smoker status. Suppose once again that a life insurance agent is deciding whether to approve or deny a potential customer for insurance, such that their utility matrix is shown in Table 4.1. The insurance agent can learn whether the potential customer is a Marlboro smoker, a smoker of other brands, or a non-smoker. The causal conditional probability distribution over possible mortality ages, given interventions on smoker status, is given in Table 4.2. Whether an intervention makes it the case that the customer is a Marlboro smoker or a smoker of other brands, the same action, denying life insurance, has maximum

	Death ≤ 70	Death > 70
Approve	-70.32	10
Deny	0	0

Tab. 4.5: Utility Matrix For Life Insurance Decision

	Death ≤ 70	Death > 70
Smoker	.2525	.7475
Non-Smoker	.051	.949

Tab. 4.6: Causal Conditional Probability Table

expected utility. Thus, these two smoker statuses are pragmatic causal equivalents. Further, if an intervention is performed such that the customer is a non-smoker, then approving them for life insurance has maximum expected utility. Thus, being a non-smoker is not a pragmatic causal equivalent of either of the other two smoker statuses. Next, one can check that the inequality in the definition of pragmatic effect equivalence is only satisfied for the effect values 'Death < 50' and 'Death $50 - 70$'. Thus, they are pragmatic effect equivalents.

If we coarsen according to the equivalence relations defined in the previous paragraph, we get the utility matrix and conditional probability distribution shown in Tables 4.5 and 4.6. Assuming a uniform prior probability distribution over possible interventions on the customer's smoker status, the fair price of causal information about the customer's smoker status is approximately 1.967 (calculation in appendix), and this is the maximum value for the fair price of causal information that can be achieved by coarsening the column and row variables in Tables 4.1 and 4.2, provided that we assume a prior probability distribution over the possible interventions on the row variable that is consistent with the prior distribution over the more fine-grained row variable, i.e. $p(do(Smoker)) = \frac{2}{3}$, $p(do(Non - smoker)) = \frac{1}{3}$. This shows that the proposed procedure for pragmatic causal feature learning is fit to purpose. Given a causal and probabilistic relationship between two variables, and a utility function defined over the value set of the effect variable, it identifies the coarsening such that an agent with this utility function would pay the highest possible price for information about the causal variable.

4.4.5 A Pragmatic Causal Coarsening Theorem

Recall from Section 4.3 that Chalupka et al. show that they can obtain epistemic coarsenings of the cause and effect variable from observational data, with failure only

occurring in a Lebesgue measure-zero subset of the possible probability distributions over the fine-grained variables. As discussed, this is taken to be an epistemological virtue in some segments of the literature, although the significance of results such as these for induction is a topic of controversy. Thankfully for my view, there is a similar result regarding the ability to learn the pragmatic causal coarsening of a fine-grained cause and effect variable when there is only observational data and the supposition of a causal relationship between C' and E'. Specifically, the following proposition is true:

Proposition 17 (Pragmatic Causal Coarsening Theorem). *Let C', E', and Z' be variables in a graph $\mathcal{G}'$ such that C' is an ancestor of E' and Z' is a possible confounder of the causal relationship between C' and E'. Consider the set of possible joint probability distributions over C', E', and Z'. Let C^{pc} and C^{oc} be the epistemic causal coarsening and observational causal coarsening of C', respectively, in any such probability distribution. The set of joint probability distributions over C', E', and Z' such that the value set of C^{pc} is not a quotient set of the value set of C^{oc} is Lebesgue measure zero within the set of all possible joint distributions over C', E', and Z'.*

This proposition follows as a corollary of Chalupka et al.'s causal coarsening theorem. In the proof, it is shown that the pragmatic causal coarsening C^{pc} of some causal variable C' is always a coarsening of or identical to the epistemic causal coarsening C^{ec}. Thus, if C^{ec} is a coarsening of C' in all but a Lebesgue measure zero subset of cases, then so is C^{pc}. Those who hold that it is a positive feature of an inductive method that it only fails on a Lebesgue measure-zero subset of possible probability distributions will find that pragmatic causal feature learning fares just as well as Chalupka et al.'s method when it comes to learning coarsenings of the causal variable from observational data. For those less convinced of the epistemological significance of Chalupka et al.'s result, the significance of the corollary stated in Proposition 17 should be no less convincing.

There is no formal analog of the causal coarsening theorem for pragmatic effect coarsening. However, it is still the case that we can efficiently learn the pragmatic coarsening of an effect variable when we have only an observational conditional probability distribution over the causal variable, with failure only occurring in a Lebesgue measure-zero subset of probability distributions. This is because the following proposition is true:

Proposition 18. *For any two cause-effect pairs (C', E') and (C^{pc}, E'), where C^{pc} is the pragmatic causal coarsening of C, performing pragmatic causal feature learning on either pair returns the pair (C^{pc}, E^{ec}), where E^{ec} is the pragmatic effect coarsening of E'.*

This means that we can use the following procedure to learn the pragmatic coarsening of a pair (C', E') from observational data. First, obtain the observational coarsening of C^{oc}. Next, intervene on each value of C^{oc} and calculate the expected utilities of the actions in A under each intervention to obtain the pragmatic causal coarsening C^{pc}. This procedure will only fail on a Lebesgue measure-zero subset of probability distributions over (C', E'). Finally, perform causal feature learning on (C^{pc}, E') to obtain the pragmatically coarsened cause-effect pair (C^{pc}, E^{ec}). Thus, if Chalupka et al.'s causal coarsening theorem shows that their method of causal feature learning has a good-making feature, then my pragmatic method has the same good-making feature. This is somewhat surprising, given the formal differences between the pragmatic and epistemic frameworks for causal feature learning.

4.5 Potential Counterarguments

In this section, I consider the sorts of arguments that one might give in favor of the proposition that Chalupka et al.'s methodology for coarsening causal variables is fundamentally better than pragmatic causal feature learning. Such a counterargument could object to the way that the pragmatic method treats either the causal variable or the effect variable. Given that my treatment of the causal variable has already been discussed and defended in the previous chapter, I will focus here on my proposal for coarsening the effect variable using pragmatic considerations.

Specifically, I take it that a defender of the epistemic coarsening of the effect variable could argue that the pragmatic coarsening of the effect variable only reveals something about the agent in question, rather than revealing something about the world. To illustrate, consider a micro-variable E with values that denote the directional velocity of each particle in a box of gas during a given time interval. Suppose that the agent is deliberating between two possible actions: betting that the Kelvin temperature of the gas starts with an even number, and betting that the Kelvin temperature of the gas starts with an odd number. This will lead to a peculiar coarsening of E into a binary variable representing whether or not the gas has a temperature starting with an even number. The problem with this result is that even-numbered temperature is not a scientifically meaningful category. Thus, it could be argued, pragmatic causal feature learning runs the risk of generating scientifically irrelevant macro-variables, depending on the choice of actions that an agent entertains and the utilities associated with those actions given each value of the effect variable. Such variables tell us something about the idiosyncratic whims of particular agents, but nothing about nature.

In response, I argue that when pragmatic causal feature learning produces variables that do not represent scientifically meaningful types of events, it does so because the set of actions that an agent is deliberating between is itself lacking in scientific relevance. What counts as a scientifically meaningful category is determined at least in part by the kinds of problems that we turn to science to solve. Knowing the temperature of a box of gas is relevant for all kinds of scientific, engineering, and ordinary-life problems, namely any problems that hinge in one way or another on the temperature of the gas. By contrast, few problems that we face ever depend on whether the temperature of a gas begins with an even or odd number, and thus the partition of temperatures into these kinds of categories strikes us as out of keeping with scientific practice. To sum up, while my account of pragmatic causal feature learning can output macro-variables that would not appear in scientific practice, this will only occur if we input decision problems that would also not appear in scientific practice.

However, even if pragmatic effect coarsening ends up producing a scientifically important coarsening of the effect variable, one could argue that on my view, pragmatic effect coarsening will only discover meaningful groupings of values of the fine-grained effect variable if the agent already knows which groupings are meaningful. To illustrate, consider again a micro-variable E with values that denote the directional velocity of each particle in a box of gas during a given time interval. This time, we suppose that an agent is deliberating between a set of actions that all have equal utility for any two values of E that are consistent with the box of gas having the same Kelvin temperature. This would produce a coarse-grained effect variable telling us the temperature of the gas. However, it seems that an agent whose preferences over values of E line up perfectly with a partition of the set of values of E that corresponds to the possible temperatures of the gas already has the concept of temperature in mind when they consider the fine-grained micro-variable. Thus, the coarsening of E according to temperature is not so much learned from the data as it is pre-formulated by the agent, and then imputed onto the more fine-grained data set. This calls into question whether, at least as far as the effect variable is concerned, pragmatic causal feature learning deserves to be called a learning method at all.

In response, I would argue that just as we can discover the conditional probability distribution over an effect variable, given each value of its cause, so too can we discover the utilities of various actions when they are paired with different values of an effect variable. Returning to the previous example, suppose that placing a turbine in a box of gas has the same utility, measured in terms of wattage generated, whenever the micro-variable E takes either of two values that are consistent with the box of gas having the same temperature. This empirical discovery does not imply that the agent interested in generating electricity already possesses the concept of temperature. Rather, it demonstrates that such an agent can begin to discover the

	Death < 50	Death 50-70	Death 70-90	Death > 90
Marlbro Smoker	.003	.3	.694	.003
Other Smoker	.002	.2	.796	.002
Non-Smoker	.001	.05	.948	.001

Tab. 4.7: Causal Conditional Probability Table

concept of temperature by observing that they can generate the same amount of electricity when the gas is in either of several microstates. This leads the agent to group the microstates together, thereby discovering the concept of temperature.

To be clear, this is a highly idealized rational reconstruction of how coarse-grained concepts like temperature can be pragmatically discovered. However, Chalupka et al.'s model of causal feature learning is also a highly idealized picture of macro-variable discovery, containing "a set of assumptions that do not necessarily hold in real-world settings" (2017, p. 151). As such, the charge that pragmatic causal feature learning is not really a discovery method does not stick. Like Chalupka et al.'s approach, it is an idealized model of macro-variable discovery, albeit one that rests on different formal foundations.

Additionally, Chalupka et al.'s proposal for coarsening the effect variable can sometimes deliver results that seem out of sync with scientific practice. Suppose that a person's mortality age is represented by an effect variable with the value set $\{\text{Death} < 50, \text{Death } 50-70, \text{Death } 70-90, \text{Death} > 90\}$ and that the conditional probability distribution over these possible mortality ages, given different possible smoker statuses, is given in Table 4.7. On Chalupka et al.'s model of effect coarsening, the values 'Death < 50' and 'Death > 90' should be coarsened into a single value 'Death < 50 or > 90'. However, this coarse-grained value seems not to pick out a meaningful scientific category. After all, dying before age fifty and dying after age ninety are very different outcomes from both an ontological and a pragmatic standpoint. By contrast, my account would group these two categories together only if the same action had maximal expected utility regardless of whether a person dies before fifty or after ninety, i.e. only if there is some pragmatic reason for grouping the two values together.

4.6 Conclusion

This chapter has proposed and defended a procedure for causal feature learning that rivals a procedure previously articulated by Chalupka et al. The procedure is

explicitly pragmatic in that it represents and uses agential interests in outcomes to discover a coarsening of a fine-grained causal relationship that preserves all and only the information that is of prudential value to a given agent. The pragmatic procedure has been shown to compare favorably to Chalupka et al.'s procedure in some respects. Importantly, this chapter has been wholly theoretical and not concerned with the actual implementation of pragmatic causal feature learning. As such, an avenue for future work would be to explore how Chalupka et al.'s algorithms for causal feature learning can be augmented in order to perform pragmatic causal feature learning on actual data.

4.7 Appendix

4.7.1 Calculations

Fair Price of Causal Information for Tables 4.1 and 4.2

If there is an intervention such that the customer is a Marlboro smoker, the action with maximum expected utility is Deny, which has an expected utility of zero. If there is an intervention such that the customer is a smoker of other brands, the action with maximum expected utility is also Deny, which has an expected utility of zero. If there is an intervention such that the customer is a non-smoker, the action with maximum expected utility is Approve, which has an expected utility of $.001(-100) + .05(-70) + .949(10) = 5.89$. Assuming that all three interventions have equal prior probability, the expected utility of learning which intervention has been performed on the customer's smoker status is $\frac{1}{3}(0) + \frac{1}{3}(0) + \frac{1}{3}(5.89) = 1.96\bar{3}$. If the agent does not learn which intervention has been performed, the action with maximal expected utility is Deny, which has an expected utility of 0. Thus, the fair price of causal information is $1.96\bar{3} - 0 = 1.96\bar{3}$.

Coarsenings to Produce Tables 4.3 and 4.4

The probability of dying at 70 or younger if one is a Marlboro smoker is the sum of the probability of dying before 50 if one is a Marlboro smoker and the probability of dying between 50 and 70 if one is a Marlboro smoker, i.e. $.003 + .3 = .303$. The probability of dying at 70 or younger if one is a smoker of other brands is the sum of the probability of dying before 50 if one is a smoker of other brands and the probability of dying between 50 and 70 if one is a smoker of other brands, i.e. $.002 + .2 = .202$. The probability of dying at 70 or younger if one is a non-smoker

is the sum of the probability of dying before 50 if one is a non-smoker and the probability of dying between 50 and 70 if one is a non-smoker, i.e. $.001 + .05 = .051$. In this scenario, the marginal probability that a person dies before 50 is $.002$, and the marginal probability that a person dies between 50 and 70 is $.183$. This means that the utility associated with approving life insurance when the patient dies before 70 is $\frac{.002}{.185}(-100) + \frac{.183}{.185}(-70) = -70.32$.

Fair Price of Causal Information for Tables 4.3 and 4.4

If there is an intervention such that the customer is a Marlboro smoker, the action with maximum expected utility is Deny, which has an expected utility of zero. If there is an intervention such that the customer is a smoker of other brands, the action with maximum expected utility is also Deny, which has an expected utility of zero. If there is an intervention such that the customer is a non-smoker, the action with maximum expected utility is Approve, which has an expected utility of $.051(-70.32) + .949(10) \approx 5.90$. Assuming that all three interventions have equal prior probability, the expected utility of learning which intervention has been performed on the customer's smoker status is $\frac{1}{3}(0) + \frac{1}{3}(0) + \frac{1}{3}(5.90) \approx 1.967$. If the agent does not learn which intervention has been performed, the action with maximal expected utility is Deny, which has an expected utility of 0. Thus, the fair price of causal information is approximately $1.967 - 0 = 1.967$.

Fair Price of Causal Information for Tables 4.5 and 4.6

If there is an intervention such that the customer is a Marlboro smoker, the action with maximum expected utility is Deny, which has an expected utility of zero. If there is an intervention such that the customer is a non-smoker, the action with maximum expected utility is Approve, which has an expected utility of $.051(-70.32) + .949(10) \approx 5.90$. The expected utility of learning which intervention has been performed on the customer's smoker status is $\frac{2}{3}(0) + \frac{1}{3}(5.90) \approx 1.967$. If the agent does not learn which intervention has been performed, the action with maximal expected utility is Deny, which has an expected utility of 0. Thus, the fair price of causal information is approximately $1.967 - 0 = 1.967$.

4.7.2 Proof of Proposition 10

Proof. See Chalupka et al. (2016b), Supplementary Material C. □

4.7.3 Proof of Proposition 11

For this proof and several subsequent proofs, it will be helpful to have in mind the following equation for the fair price of causal information for a causal variable C with m values and effect variable E with n values:

$$\pi(u(\cdot), A, E, C) = \sum_{j=1}^{m} p(do(c_j)) \sum_{i=1}^{n} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_j)), e_i) p(e'_s|do(c_j)) - max_{a_k \in A} eu(a_k) \quad (4.4)$$

Next, let us prove the following lemma, which will shorten several of the subsequent proofs.

Lemma 2. *If the value sets of C and E are quotient sets of the value sets of C' and E', respectively, then $\max_{a_k \in A} eu(a_k)$ for the variable pair (C', E') equals $\max_{a_k \in A} eu(a_k)$ for the variable pair (C, E).*

Proof. Begin with the fine-grained variable pair (C', E'). One can define $max_{a_k \in A} eu(a_k)$ as follows, when E' has w values:

$$\max_{a_k \in A} eu(a_k) = \max_{a_k \in A} \sum_{s=1}^{w} p(e'_s) u(a_k, e'_s) \quad (4.5)$$

Any value e_i of the coarse-grained variable E can be re-written as follows: $e_i = \{e'_{\alpha i}, e'_{\alpha i+1}, ..., e'_{\alpha i+\gamma i}\}$. This allows us to re-write the value set of E' as follows:

$$R_{E'} = \{e'_{\alpha 1}, e'_{\alpha 1+1}, ..., e'_{\alpha 1+\gamma 1}, e'_{\alpha 2}, e'_{\alpha 2+1}, ..., e'_{\alpha 2+\gamma 2}, ..., e'_{\alpha n}, e'_{\alpha n+1}, ..., e'_{\alpha n+\gamma n}\} \quad (4.6)$$

This representation of the value set of E', in addition to the supposition that the value set of E is a quotient set of the value set of E', allows us to re-write the equation (4.5), assuming that E has n values:

$$\max_{a_k \in A} eu(a_k) = \max_{a_k \in A} \sum_{i=1}^{n} \sum_{\tau=0}^{\gamma n} p(e'_{\alpha n+\tau}) u(a_k, e'_{\alpha n+\tau}) \quad (4.7)$$

Next, note that for each value e_i, the following two equations hold:

$$p(e_i) = \sum_{\tau=0}^{\gamma n} p(e'_{\alpha n+\tau}) \quad (4.8)$$

$$u(a_k, e_i) = \sum_{\tau=0}^{\gamma n} u(a_k, e'_{\alpha n+\tau}) \frac{p(e'_{\alpha n+\tau})}{p(e_i)} \text{ for all } a_k. \quad (4.9)$$

This allows us to re-write (4.7) as follows:

$$\max_{a_k \in A} eu(a_k) = \max_{a_k \in A} \sum_{i=1}^{n} \sum_{\tau=0}^{\gamma n} p(e'_{\alpha n+\tau}) u(a_k, e'_{\alpha n+\tau}) = \max_{a_k \in A} \sum_{i=1}^{n} p(e_i) u(a_k, e_i) \quad (4.10)$$

Thus, coarsening or refining the effect variable does not change the value of $\max_{a_k \in A} eu(a_k)$. To see that coarsening the causal variable does not change the value of $\max_{a_k \in A} eu(a_k)$, note that the law of total probability implies the following, for a causal variable C' with q values:

$$\max_{a_k \in A} \sum_{i=1}^{n} p(e_i) u(a_k, e_i) = \max_{a_k \in A} \sum_{i=1}^{n} \sum_{l=1}^{q} p(e_i | do(c'_l)) p(do(c'_l)) u(a_k, e_i) \quad (4.11)$$

Using similar logic as in the previous step, we can re-write this equation as follows, for a coarse-grained causal variable C with m values:

$$\begin{aligned} \max_{a_k \in A} \sum_{i=1}^{n} p(e_i) u(a_k, e_i) &= \max_{a_k \in A} \sum_{i=1}^{n} \sum_{l=1}^{q} p(e_i | do(c'_t)) p(do(c'_t)) u(a_k, e_i) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{\tau=0}^{\gamma j} p(e_i | do(c'_{\alpha j+\tau})) p(do(c'_{\alpha j+\tau})) u(a_k, e_i) \quad (4.12) \end{aligned}$$

This entails that:

$$\max_{a_k \in A} \sum_{i=1}^{n} p(e_i) u(a_k, e_i) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(e_i | do(c_j)) p(do(c_j)) u(a_k, e_i) \quad (4.13)$$

Thus, coarsening or refining the effect variable does not change the value of $\max_{a_k \in A} eu(a_k)$. □

This lemma allows us to ignore the term $\max_{a_k \in A} eu(a_k)$ when comparing the fair price of causal information when comparing the coarsenings and refinements of the variables involved. We turn now to the proof of Proposition 11:

Proof. Let us proceed by contrapositive proof. Suppose that $\pi(u(\cdot), A, E', C^{pc}) \neq \pi(u(\cdot), A, E', C')$. This, along with Lemma 2, implies the following inequality, where C^{pc} has y values:

$$\begin{aligned} &\sum_{h=1}^{y} p(do(c_h^{pc})) \sum_{s=1}^{w} u(\operatorname*{argmax}_{a_k \in A} eu(a_k | do(c_h^{pc})), e'_s) p(e'_s | do(c_h^{pc})) \\ &\qquad \neq \sum_{l=1}^{q} p(do(c'_l)) \sum_{s=1}^{w} u(\operatorname*{argmax}_{a_k \in A} eu(a_k | do(c'_l)), e'_s) p(e'_s | do(c'_l)) \quad (4.14) \end{aligned}$$

This inequality implies in turn that for some value c_h^{pc}, and set of values $\{c'_1, c'_2, ..., c'_\alpha\}$ such that $c_h^{pc} = \{c'_1, c'_2, ..., c'_\alpha\}$, the following inequality holds:

$$p(do(c_h^{pc})) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_h^{pc})), e'_s) p(e'_s|do(c_h^{pc})) \neq \sum_{l=1}^{\alpha} p(do(c'_l)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e'_s) p(e'_s|do(c'_l)) \quad (4.15)$$

Since $c_h^{pc} = \{c'_1, c'_2, ..., c'_\alpha\}$, $p(e'_s|do(c_h^{pc})) = \sum_{j=1}^{\alpha} p(e'_s|do(c'_l)) \frac{p(do(c'_l))}{p(do(c_h^{pc}))}$ for any e'_s. Thus, we can re-write (4.15) as follows:

$$p(do(c_h^{pc})) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_h^{pc})), e'_s) \sum_{l=1}^{\alpha} p(e'_s|do(c'_l)) \frac{p(do(c'_l))}{p(do(c_h^{pc}))} \neq \sum_{j=1}^{\alpha} p(do(c'_l)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e'_s) p(e'_s|do(c'_l)) \quad (4.16)$$

Which reduces to:

$$\sum_{l=1}^{\alpha} p(do(c'_l)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_h^{pc})), e'_s) p(e'_s|do(c'_l)) \neq \sum_{l=1}^{\alpha} p(do(c'_l)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e'_s) p(e'_s|do(c'_l)) \quad (4.17)$$

The truth of this inequality would entail the existence of a value c'_l of the variable C' such that $\text{argmax}_{a_k \in A}\, eu(a_k|do(c_h^{pc})) \neq \text{argmax}_{a_k \in A}\, eu(a_k|do(c'_l))$. But this would mean that there are elements c'_l and c'_t in the equivalence class $c_h^{pc} = \{c'_1, c'_2, ..., c'_\alpha\}$ such that $\text{argmax}_{a_k \in A}\, eu(a_k|do(c'_l)) \neq \text{argmax}_{a_k \in A}\, eu(a_k|do(c'_t))$. This would imply that C^{pc} is not the pragmatic coarsening of C', which proves the proposition. □

4.7.4 Proof of Proposition 12

Proof. Partition the value set R_C into the sets R_C^1 and R_C^2. Partition the value set $R_{C^{pc}}$ into the sets $R_{C^{pc}}^1$ and $R_{C^{pc}}^2$. Design these sets such that there is a function $\phi(\cdot) : R_C^1 \rightarrow R_{C^{pc}}^1$ such that $\phi(c_j) = c_h^{pc}$ if and only if $c_j = c_h^{pc}$ or $c_j \subseteq c_h^{pc}$.[3] By contrast,

[3] Recall from the introduction that because C and C^{pc} are both coarsenings of C' values of C and C^{pc}, can stand in the subset-superset relation to one another.

for all $c_j \in R^2_C$ and all $c^{pc}_h \in R^2_{C^{pc}}$, $c_j \neq c^{pc}_h$ and $c_j \not\subseteq c^{pc}_h$. If $R^1_C = \{c_1, c_2, ..., c_\alpha\}$ and $R^1_{C^{pc}} = \{c^{pc}_1, c^{pc}_2, ..., c^{pc}_\beta\}$ then the following holds:

$$\sum_{j=1}^{\alpha} p(do(c_j)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_j)), e'_s) p(e'_s|do(c_j))$$
$$= \sum_{h=1}^{\beta} p(do(c^{pc}_h)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c^{pc})), e'_s) p(e'_s|do(c^{pc}_h)) \quad (4.18)$$

For any $c_j \in R^2_C$ such that $c_j = \{c'_1, c'_2, ..., c'_\eta\} \subseteq R^2_{C'}$, the following equation holds:

$$p(do(c_j)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_j)), e'_s) p(e'_s|do(c_j))$$
$$= p(do(c_j)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_j)), e'_s) \sum_{l=1}^{\eta} p(e'_s|do(c'_l)) \frac{p(do(c'_l))}{p(do(c_j))}$$
$$= \sum_{l=1}^{\eta} p(do(c'_l)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_j)), e'_s) p(e'_s|do(c'_l)) \quad (4.19)$$

The fact that there is no value of C^{pc} such that $c_j = c^{pc}_h$ or $c_j \subseteq c^{pc}_h$ entails that some values of C' in c_j do not stand in the relation $\sim_{pc}$ to each other, which entails that for there must be at least one $c'_l \in c_j$ such that $\operatorname{argmax}_{a_k \in A} eu(a_k|do(c_j)) \neq \operatorname{argmax}_{a_k \in A} eu(a_k|do(c'_l))$. For any such pair (c'_l, c_j), the following inequality holds:

$$\sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_j)), e'_s) p(e'_s|do(c'_l))$$
$$< \sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c'_l)), e'_s) p(e'_s|do(c'_l)) \quad (4.20)$$

To deny this inequality would lead to a contradiction in terms; the expected utility of the action with maximum expected utility when $C' = c'_l$ would not be maximal when $C' = c'_l$. For any $c'_l \in c_j$ such that $\operatorname{argmax}_{a_k \in A} eu(a_k|do(c_j)) = \operatorname{argmax}_{a_k \in A} eu(a_k|do(c'_l))$, the following holds:

$$\sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_j)), e'_s) p(e'_s|do(c'_l))$$
$$= \sum_{s=1}^{w} u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c'_l)), e'_s) p(e'_s|do(c'_l)) \quad (4.21)$$

This all implies the following:

$$\sum_{l=1}^{\eta} p(do(c_l')) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_j)), e_s')p(e_s'|do(c_l')) \\ < \sum_{l=1}^{\eta} p(do(c_l')) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_l')), e_s')p(e_s'|do(c_l')) \quad (4.22)$$

Together, (4.19) and (4.22) imply the following:

$$p(do(c_j)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_j)), e_s')p(e_s'|do(c_j)) \\ < \sum_{l=1}^{\eta} p(do(c_l')) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_l')), e_s')p(e_s'|do(c_l')) \quad (4.23)$$

Repeating this process for every value of C, we get the result that for a variable C with m values, and variable C' with q values, if R^2_C and $R^2_{C^{pc}}$ are non-empty, then the following holds:

$$\sum_{j=1}^{m} p(do(c_j)) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_j)), e_s')p(e_s'|do(c_j)) \\ < \sum_{l=1}^{q} p(do(c_l')) \sum_{s=1}^{w} u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c_l')), e_s')p(e_s'|do(c_l')) \quad (4.24)$$

This, along with Lemma 2, implies that if if R^2_C and $R^2_{C^{pc}}$ are non-empty, then $\pi(u(\cdot), A, E', C) < \pi(u(\cdot), A, E', C')$. Proposition 11 implies that $\pi(u(\cdot), A, E', C') = \pi(u(\cdot), A, E', C^{pc})$. Thus, if R^2_C and $R^2_{C^{pc}}$ are non-empty, then $\pi(u(\cdot), A, E', C) < \pi(u(\cdot), A, E', C^{pc})$. If $\pi(u(\cdot), A, E', C) = \pi(u(\cdot), A, E', C^{pc})$, then (4.20) does not hold for any values c_l' and c_j. This implies that R^2_C and $R^2_{C^{pc}}$ are empty, which implies that the value set of C^{pc} is a quotient set of the value set of C. Thus, $\pi(u(\cdot), A, E', C) \leq \pi(u(\cdot), A, E', C^{pc})$, and if $\pi(u(\cdot), A, E', C) = \pi(u(\cdot), A, E', C^{pc})$, then the value set of C^{pc} is a quotient set of the value set of C. □

4.7.5 Proof of Proposition 13

Proof. Let $\{e_1', e_2', ..., e_\alpha'\}$ be a set of values of a fine-grained effect variable E'. Let e_i be a value of a coarse-grained effect variable E such that $E = e_i$ if and only if

E' takes a value in the set $\{e'_1, e'_2, ..., e'_\alpha\}$. Suppose that for all such e_i and all c'_l, the following holds, assuming that C' has q values:

$$\sum_{l=1}^{q} p(do(c'_l))p(e_i|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e_i) \geq \sum_{l=1}^{q} p(do(c'_l)) \sum_{l=1}^{\alpha} p(e'_s|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e'_s) \quad (4.25)$$

Since the above inequality holds for all e_i and all c'_l, the following also holds, under the assumption that E has n values and E' has w values:

$$\sum_{l=1}^{q} p(do(c'_l)) \sum_{i=1}^{n} p(e_i|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e_i) \geq \sum_{l=1}^{q} p(do(c'_l)) \sum_{s=1}^{w} p(e'_u|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e'_u) \quad (4.26)$$

This, along with Lemma 2, entails that $\pi(u(\cdot), A, E, C') \geq \pi(u(\cdot), A, E', C')$. □

4.7.6 Proof of Proposition 14

Proof. This proof proceeds similarly to the proof of Proposition 12. Partition the value set R_E into the sets R^1_E and R^2_E. Partition the value set $R_{E^{pe}}$ into the sets $R^1_{E^{pe}}$ and $R^2_{E^{pe}}$. Design these sets such that there is a function $\phi(\cdot) : R^1_E \to R^1_{E^{pe}}$ such that $\phi(e_i) = e^{pe}_v$ if and only if $e_i = e^{pe}_v$ or $e_i \subseteq e^{pe}_v$. By contrast, for all $e_i \in R^2_E$ and all $e^{pe}_v \in R^2_{E^{pe}}$, $e_i \neq e^{pe}_v$ and $e_i \not\subseteq e^{pe}_v$. If $R^1_E = \{e_1, e_2, ..., e_\alpha\}$ and $R^1_{E^{pe}} = \{e^{pe}_1, e^{pe}_2, ..., e^{pe}_\beta\}$ then the following holds:

$$\sum_{l=1}^{q} p(do(c'_l)) \sum_{i=1}^{\alpha} p(e_i|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e_i) \leq \sum_{l=1}^{q} p(do(c'_l)) \sum_{v=1}^{\beta} p(e^{pe}_v|do(c'_l))u(\underset{a_k \in A}{\text{argmax}}\, eu(a_k|do(c'_l)), e^{pe}_v) \quad (4.27)$$

Next, consider $R^2_E = \{e_{\alpha+1}, e_{\alpha+2}, ..., e_{\alpha+\gamma}\}$ and $R^2_{E^{pe}} = \{e^{pe}_{\beta+1}, e^{pe}_{\beta+2}, ..., e^{pe}_{\beta+\delta}\}$. Both of these sets are partitions of the same subset $R^2_{E'}$ of the value set of a fine-grained variable E', where $R^2_{E'} = \{e'_{\epsilon+1}, e'_{\epsilon+2}, ..., e'_{\epsilon+\zeta}\}$. For any $e_i \in R^2_E$ such that $e_i = \{e'_1, e'_2, ..., e'_\eta\} \subseteq R^2_{E'}$, the fact that there is no value of E^{pe} such that $e_i = e^{pe}_v$ or

$e_i \subseteq e_v^{pe}$ entails that some values of E' in e_i do not stand in the relation $\sim_{pe}$ to each other, which entails that the following inequality holds:

$$\sum_{l=1}^{q} p(do(c_l'))p(e_i|do(c_l'))u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_l')), e_i) < \sum_{l=1}^{q} p(do(c_l')) \sum_{s=1}^{\eta} p(e_s'|do(c_l'))u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_l')), e_s') \quad (4.28)$$

Given that this holds for all $e_i \in R_E^2$, and given that R_E^2 is a partition of $R_{E'}^2$, the following also holds:

$$\sum_{l=1}^{q} p(do(c_l')) \sum_{\tau=1}^{\gamma} p(e_{\alpha+\tau}|do(c_l'))u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_l')), e_{\alpha+\tau}) < \sum_{l=1}^{q} p(do(c_l')) \sum_{\xi=1}^{\zeta} p(e_{\epsilon+\xi}'|do(c_l'))u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_l')), e_{\epsilon+\xi}') \quad (4.29)$$

By contrast, for any $e_v^{pe} \in R_{E^{pe}}^2$ such that $e_v^{pe} = \{e_1', e_2', ..., e_\theta'\} \subseteq R_{E'}^2$, all values of E' in e_v^{pe} stand in the relation $\sim_{pe}$ to each other, which entails that the following inequality holds:

$$\sum_{l=1}^{q} p(do(c_l'))p(e_v^{pe}|do(c_l'))u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_l')), e_v^{pe}) \geq \sum_{l=1}^{q} p(do(c_l')) \sum_{s=1}^{\theta} p(e_s'|do(c_l'))u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_l')), e_s') \quad (4.30)$$

Given that this holds for all $e_v^{pe} \in R_{E^{pe}}^2$, and given that $R_{E^{pe}}^2$ is a partition of $R_{E'}^2$, the following also holds:

$$\sum_{l=1}^{q} p(do(c_l')) \sum_{\tau=1}^{\delta} p(e_{\beta+\tau}^{pe}|do(c_l'))u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_l')), e_{\beta+\tau}^{pe}) \geq \sum_{l=1}^{q} p(do(c_l')) \sum_{\xi=1}^{\zeta} p(e_{\epsilon+\xi}'|do(c_l'))u(\underset{a_k \in A}{\operatorname{argmax}}\, eu(a_k|do(c_l')), e_{\epsilon+\xi}') \quad (4.31)$$

Let n be the number of values in E and let z be the number of values in E^{pe}. Since $\{R_E^1, R_E^2\}$ partitions the value space of E and $\{R_{E^{pe}}^1, R_{E^{pe}}^2\}$ partitions the value space of E^{pe}, the following equations hold:

$$\begin{aligned}
&\sum_{l=1}^{q} p(do(c_l')) \sum_{i=1}^{n} p(e_i|do(c_l'))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_l')), e_i) \\
&= \sum_{l=1}^{q} p(do(c_l')) \sum_{i=1}^{\alpha} p(e_i|do(c_l'))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_l')), e_i) \\
&\quad + \sum_{l=1}^{q} p(do(c_l')) \sum_{\tau=1}^{\gamma} p(e_{\alpha+\tau}|do(c_l'))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_l')), e_{\alpha+\tau}) \qquad (4.32)
\end{aligned}$$

$$\begin{aligned}
&\sum_{l=1}^{q} p(do(c_l')) \sum_{v=1}^{z} p(e_v^{pe}|do(c_l'))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_l')), e_v^{pe}) \\
&= \sum_{l=1}^{q} p(do(c_l')) \sum_{v=1}^{\beta} p(e_v^{pe}|do(c_l'))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_l')), e_v^{pe}) \\
&\quad + \sum_{l=1}^{q} p(do(c_l')) \sum_{\tau=1}^{\delta} p(e_{\beta+\tau}^{pe}|do(c_l'))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_l')), e_{\beta+\tau}^{pe}) \qquad (4.33)
\end{aligned}$$

If R_E^2 and $R_{E^{pe}}^2$ are non-empty, then Equations (4.27), (4.29), and (4.31)-(4.33) imply the following:

$$\begin{aligned}
&\sum_{l=1}^{q} p(do(c_l')) \sum_{v=1}^{z} p(e_v^{pe}|do(c_l'))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_l')), e_v^{pe}) \\
&\qquad > \sum_{l=1}^{q} p(do(c_l')) \sum_{i=1}^{n} p(e_i|do(c_l'))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_l')), e_i) \qquad (4.34)
\end{aligned}$$

Along with Lemma 2, this implies that $\pi(u(\cdot), A, E, C') \leq \pi(u(\cdot), A, E^{pe}, C')$. Further, if $\pi(u(\cdot), A, E, C') = \pi(u(\cdot), A, E^{pe}, C')$, then (4.28) does not hold for any pair (e_s', e_i), which implies that R_E^2 and $R_{E^{pe}}^2$ are empty. This implies in turn that there are no values of E and E^{pe} such that $e_i \neq e_v^{pe}$ and $e_i \not\subseteq e_v^{pe}$, which implies that the value set of E^{pe} is a quotient set of the value set of E. Thus, $\pi(u(\cdot), A, E, C') \leq \pi(u(\cdot), A, E^{pe}, C')$, and if $\pi(u(\cdot), A, E, C') = \pi(u(\cdot), A, E^{pe}, C')$, then the value set of E^{pe} is a quotient set of the value set of E. □

4.7.7 Proof of Proposition 15

Proof. If C^{pc} is the pragmatic causal coarsening of C', and the value set of C is a quotient set of the value set of C', then Proposition 12 entails that $\pi(u(\cdot), A, E, C) \leq \pi(u(\cdot), A, E, C^{pc})$ and if $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E, C^{pc})$, then the value set of C^{pc} is a quotient set of the value set C. If E^{pe} is the pragmatic effect coars-

ening of E' and E is a quotient set of the value set of E', then Proposition 14 entails that $\pi(u(\cdot), A, E, C^{pc}) \leq \pi(u(\cdot), A, E^{pc}, C^{pc})$ and if $\pi(u(\cdot), A, E, C^{pc}) = \pi(u(\cdot), A, E^{pe}, C^{pc})$, then the value set of E^{pe} is a quotient set of the value set E. Thus, if all antecedent conditions hold, then $\pi(u(\cdot), A, E, C) \leq \pi(u(\cdot), A, E^{pe}, C^{pc})$. Further, if $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E^{pe}, C^{pc})$, given that Proposition 11 says that $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E, C^{pc})$, it follows via the transitivity of identity that $\pi(u(\cdot), A, E, C^{pc}) = \pi(u(\cdot), A, E^{pe}, C^{pc})$, which entails via Proposition 14 that the value set of E^{pe} is a quotient set of the value set of E. If $\pi(u(\cdot), A, E, C^{pc}) = \pi(u(\cdot), A, E^{pe}, C^{pc})$ and $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E^{pe}, C^{pc})$, it follows via the transitivity of identity that $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E, C^{pc})$, which entails via Proposition 12 that the value set of C^{pc} is a quotient set of the value set of C. Thus, if $\pi(u(\cdot), A, E, C) = \pi(u(\cdot), A, E^{pe}, C^{pc})$, then the value set of C^{pc} is a quotient set of the value set of C and the value set of E^{pe} is a quotient set of the value set E. □

4.7.8 Proof of Proposition 16

Proof. If C^{pc} were not the unique pragmatic causal coarsening of C' with respect to E', or E^{pe} were not the unique pragmatic effect coarsening of E' with respect to C', then either: 1) there would be a pair of values c'_l and c'_t such that $c'_l \sim_{pc} c'_t$ with respect to E' and such that it is not the case that $c'_l \sim_{pc} c'_t$ with respect to E' 2) there would be a pair of values e'_s and e'_u such that $e'_s \sim_{pe} e'_u$ with respect to C' and such that it is not the case that $e'_s \sim_{pe} e'_u$ with respect to C' , according to a single utility function and probability distribution. Since both of these disjuncts is a contradiction, C^{pc} must be the unique pragmatic causal coarsening and of C' with respect to E' and E^{pe} must be the unique pragmatic effect coarsening of E' with respect to C'.

To show that C^{pc} is the unique pragmatic causal coarsening of itself with respect to E^{pe}, and that E^{pe} is the unique pragmatic causal coarsening of itself with respect to C^{pc}, let $C^{\dagger}$ and $E^{\dagger}$ be any pragmatic causal coarsening and pragmatic effect coarsening of C^{pc} and E^{pe}, respectively. Since $C^{\dagger}$ and $E^{\dagger}$ are the pragmatic causal and effect coarsening of C^{pc} and E^{pe}, and since the value sets of C^{pc} and E^{pe} are quotient sets of themselves, Proposition 15 implies that $\pi(u(\cdot), A, E^{pe}, C^{pc}) \leq \pi(u(\cdot), A, E^{\dagger}, C^{\dagger})$. Since the value sets of $C^{\dagger}$ and $E^{\dagger}$ are quotient sets of the value sets of C' and E', respectively, Proposition 15 implies that $\pi(u(\cdot), A, E^{\dagger}, C^{\dagger}) \leq \pi(u(\cdot), A, E^{pe}, C^{pc})$. So $\pi(u(\cdot), A, E^{\dagger}, C^{\dagger}) = \pi(u(\cdot), A, E^{pe}, C^{pc})$. This implies, also via proposition 15, that the value sets of C^{pc} and E^{pe} are quotient sets of the value sets of $C^{\dagger}$ and $E^{\dagger}$, respectively. Since $C^{\dagger}$ and $E^{\dagger}$ are, by assumption, the pragmatic causal coarsening and pragmatic effect coarsening of C^{pc} and E^{pe}, respectively, it follows that the value sets of $C^{\dagger}$ and $E^{\dagger}$ are quotient sets on the value sets of C^{pc} and E^{pe}, respectively. If the value sets of two variables, defined over the same

outcome space, are quotient sets of each other, then they are the same variable. Thus, $C^{pc} = C^{\dagger}$ and $E^{pe} = E^{\dagger}$, proving the proposition. □

4.7.9 Proof of Proposition 17

Proof. We first show that for a given joint probability distribution $p(\cdot)$, utility function $u(\cdot)$, set of actions A, and effect variable E', the value set of C^{pc} is a quotient set of the value set of C^{ec} according to the equivalence relation $\sim_{pc}$. To see this, note that if the value set of C^{pc} were not a quotient set of the value set of C^{ec} according to the equivalence relation $\sim_{pc}$, then there would be two values (c'_l, c'_t) of the fine-grained variable C' such that $c'_l \sim_{ec} c'_t$ and such that it is not the case that $c'_l \sim_{pc} c'_t$. This means that for all e'_s, $p(e'_s|do(c'_l)) = p(e'_s|do(c'_l))$ but that there is an action a_k such that $\operatorname{argmax}_{a_k \in A} eu(a_k|do(c'_l)) \neq \operatorname{argmax}_{a_k \in A} eu(a_k|do(c'_t))$. This second condition can be expanded as follows:

$$\operatorname*{argmax}_{a_k \in A} \sum_{i=1}^{n} p(e_i|do(c'_l))u(a_k, e'_s) \neq \sum_{i=1}^{n} p(e_i|do(c'_t))u(a_k, e'_s) \tag{4.35}$$

This inequality cannot hold for a fixed utility function if for all e'_s, $p(e'_s|do(c'_l)) = p(e'_s|do(c'_l))$. Thus, it must be the case that the value set of C^{pc} is a quotient set of the value set of C^{ec}. Recall that Proposition 10 states that only on a Lebesgue measure-zero subset of possible probability distributions is the value set of C^{ec} not a quotient set of the value set of C^{oc}. Since the value set of C^{pc} is always a quotient set of the value set of C^{ec}, it follows that for a fixed utility function $u(\cdot)$, set of actions A, and effect variable E', the value set of C^{pc} is a quotient set of the value set of C^{oc} in all but a Lebesgue measure-zero subset of possible probability distributions. □

4.7.10 Proof of Proposition 18

Proof. By definition, performing pragmatic causal feature learning on the pair (C', E') returns the pair (C^{pc}, E^{pe}). So we only need to show that performing pragmatic causal feature learning on the pair (C^{pc}, E') returns the same pair (C^{pc}, E^{pe}). Suppose for *reductio* that it were not the case that performing pragmatic causal feature learning on the pair (C^{pc}, E') returned the pair (C^{pc}, E^{pe}). This would require

that there is a pair of values e'_s and e'_u such that, for all c'_l and all c_l^{pe}, one of two disjuncts is true. The first disjunct says that the following two inequalities hold:

$$\begin{aligned}
&\sum_{l=1}^{q} p(do(c'_l))p(e'_s \vee e'_u|do(c'_l))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e'_s \vee e'_u) \\
&\geq \sum_{l=1}^{q} p(do(c'_l))\Big(p(e'_s|do(c'_l))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e'_s) \\
&\qquad\qquad + p(e'_s|do(c'_l))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e'_u)\Big) \qquad (4.36)
\end{aligned}$$

$$\begin{aligned}
&\sum_{l=1}^{q} p(do(c_h^{pc}))p(e'_s \vee e'_u|do(c_h^{pc}))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_h^{pc})), e'_s \vee e'_u) \\
&< \sum_{l=1}^{q} p(do(c_h^{pc}))\Big(p(e'_s|do(c_h^{pc}))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_h^{pc})), e'_s) \\
&\qquad\qquad + p(e'_s|do(c_h^{pc}))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_h^{pc})), e'_u)\Big) \qquad (4.37)
\end{aligned}$$

While the second disjunct says that the following two inequalities hold:

$$\begin{aligned}
&\sum_{l=1}^{q} p(do(c'_l))p(e'_s \vee e'_u|do(c'_l))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e'_s \vee e'_u) \\
&< \sum_{l=1}^{q} p(do(c'_l))\Big(p(e'_s|do(c'_l))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e'_s) \\
&\qquad\qquad + p(e'_s|do(c'_l))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c'_l)), e'_u)\Big) \qquad (4.38)
\end{aligned}$$

$$\begin{aligned}
&\sum_{l=1}^{q} p(do(c_h^{pc}))p(e'_s \vee e'_u|do(c_h^{pc}))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_h^{pc})), e'_s \vee e'_u) \\
&\geq \sum_{l=1}^{q} p(do(c_h^{pc}))\Big(p(e'_s|do(c_h^{pc}))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_h^{pc})), e'_s) \\
&\qquad\qquad + p(e'_s|do(c_h^{pc}))u(\operatorname*{argmax}_{a_k \in A} eu(a_k|do(c_h^{pc})), e'_u)\Big) \qquad (4.39)
\end{aligned}$$

However, when $c'_l \in c_h^{pc}$, neither of these pairs of inequalities can simultaneously be true, because $\operatorname{argmax}_{a_k \in A} eu(a_k|do(c'_l)) = \operatorname{argmax}_{a_k \in A} eu(a_k|do(c_h^{pc}))$. Thus, it must be the case that performing pragmatic causal feature learning on the pair (C^{pc}, E') returns the pair (C^{pc}, E^{pe}). □

5 Blocking the Argument for Emergent Chances

5.1 Introduction

Throughout this dissertation, it has been implicit in my argument that coarse-grained, probabilistic explanations play an important role in the special sciences. Many authors in philosophy of science—including Loewer (2001), Cohen and Callender (2009), Hoefer (2007), Sober (2010), Glynn (2010), and List and Pivato (2015b)—argue that the probabilities used in these explanations at least sometimes deserve the title of *objective chances*. As is standard, the term 'objective chance' refers here to a non-extreme probability that represents an observer-independent uncertainty about whether some event will occur. Objective chances stand in contrast to subjective or epistemic probabilities, which represent some agent's uncertainty about whether a given event will occur. Crucially, and most controversially, the authors cited above hold that special science probabilities can be objective whether or not the true fundamental physical theory of the world is deterministic or indeterministic. That is, special science objective chances can *emerge* out of a deterministic physical theory. Let us call this position *the emergent chance thesis*.

In this chapter, I focus my attention on List and Pivato's version of the emergent chance thesis, as it is the most formally precise version of the argument. I show that their argument depends on a particular formal account of *coarsening*, where coarsening is the process through which more fine-grained descriptions of events are mapped to more coarse-grained descriptions of those same events. In light of this finding, I put forward an alternative account of coarsening that blocks List and Pivato's argument for the emergent chance thesis. I argue that my account has a key advantage over List and Pivato's: unlike their account, mine allows us to articulate crucial aspects of both Kolodny and MacFarlane's (2010) "Miners Puzzle" and the statistical phenomenon known as "Simpson's Paradox".

Here is the plan for the remainder of this chapter. In Section 5.2, I give a perspicuous reconstruction of List and Pivato's formal definitions of events, probabilities, and coarsening. This provides the necessary background for Section 5.3, where I present their argument for the emergent chance thesis. In Section 5.4, I present

my rival account of coarsening, show how it blocks List and Pivato's argument for the emergent chance thesis, and demonstrate how it is better equipped to aid our reasoning with respect to the Miners Puzzle and in cases of Simpson's Paradox. In Section 5.5, I consider and respond to counterarguments to my view. In Section 5.6, I connect my arguments in this chapter to the formal approach that I take to coarse-graining in previous chapters. In Section 5.7, I offer concluding remarks.

5.2 Background on Probability and Coarsening

5.2.1 Events as Sets of Possible Worlds

Let us begin with List and Pivato's account of possible worlds. They define a *possible world* as a function $\omega(\cdot) : T \to S$, where T is a set of linearly-ordered times and S is a set of states of the world.[1] Thus, a possible world is a specification of what happens in a world at every point in time. The set of all possible worlds is denoted Ω, and the algebra $\mathcal{A}_\Omega$ is the union of the empty set and all subsets of Ω, and is closed under intersection, union, and complementation. Importantly, List and Pivato take 'the set of all possible worlds' to mean the set of all *nomologically* possible worlds, or the set of worlds compatible with some set of laws. Where Ω is the most fine-grained possible partition of the set of possible worlds, the relevant laws are taken to be the laws of physics.[2]

List and Pivato define *events* as sets of possible worlds, such that each element of the algebra $\mathcal{A}_\Omega$ is an event. To illustrate, let E be the event that Kennedy is assassinated on November 22, 1963 at 12:30pm in Dallas. In List and Pivato's framework, this means that E is the set of all possible worlds in which the time t denoting November 22, 1963 12:30pm is mapped to a state s that is consistent with Kennedy being assassinated in Dallas. Importantly, although the set E includes the actual world, it also includes other counterfactual worlds, e.g. worlds where Kennedy was assassinated by someone other than Oswald, worlds where the assassination led to a nuclear war, and worlds where, just as Kennedy was assassinated, a star in a distant galaxy went supernova. The upshot here is that events can be used to pick

[1] There are three tangential points to note here. First, List and Pivato use the term 'history' rather than 'possible world', but explicitly acknowledge that the two terms are interchangeable. Second, in a more complex version of their view, described in List and Pivato (2015a), List and Pivato define possible worlds as mappings from a set of Cartesian products of spatial and temporal locations into the set of states. Finally, although every possible world $\omega(\cdot)$ is a function, in what follows I repress the function notation $(\cdot)$ and instead just refer to a possible world as ω.

[2] Although it is tangential to my argument, List and Pivato (2015a) provide a rigorous methodology for determining the laws in a set of possible worlds, one that makes use of symmetry transformations on the set Ω.

out sets of actual and counterfactual possible worlds that share certain properties, even if those worlds are different in other respects.

5.2.2 Probabilities

A probability function $p_{\omega,t}(\cdot)$ is a mapping $\mathcal{A}_\Omega \rightarrow [0,1]$. That is, $p_{\omega,t}(\cdot)$ takes as its input a set of possible worlds and returns a value in the unit interval. Since an event is defined as a set of possible worlds, the probability function effectively assigns any event a value in the unit interval. The probability function is assumed to satisfy the Kolmogorov axioms; it assigns value 1 to the entire algebra $\mathcal{A}_\Omega$, assigns value 0 to the empty set, and assigns a probability to any countable union of disjoint events according to a principle of countable additivity.[3] As indicated by the notation, a probability function $p_{\omega,t}(\cdot)$ is indexed to a particular world and time. This is necessary because the probability assigned to the same event can change depending on the world and time at which the probability is assigned. For instance, in the actual world at times prior to his assassination, the probability that Kennedy would be assassinated in Dallas on November 22, 12:30pm was very low. However, in the actual world at times after his assassination, the probability of his being assassinated at that time and place is 1. By contrast, in worlds where he is not assassinated, the probability of Kennedy's assassination on November 22, 12:30pm is very low prior to the time of his assassination, and 0 after that time.

5.2.3 Coarsening

The crucial part of List and Pivato's formalism is their account of coarsening. They define a coarsening function $\sigma(\cdot)$, which is a mapping $S \rightarrow \mathbb{S}$, where S is a fine-grained set of states and $\mathbb{S}$ is a coarse-grained set of states. Importantly, the function need not be injective or surjective. This means that while every element of the fine-grained set of states S is mapped to some unique element of the coarse-grained set of states $\mathbb{S}$, it may be that more than one element of S is mapped to a single element of $\mathbb{S}$. This is in keeping with the claim that coarse-grained states are multiply realized by fine-grained states, meaning that multiple, distinct fine-grained states are consistent with a single coarse-grained state. Since $\sigma(\cdot)$ is a function, it is also in keeping with the claim that coarse-grained states supervene on fine-grained states; there can be no change with respect to the coarse-grained state without a corresponding change with respect to the fine-grained state.

On List and Pivato's account, when we coarsen the set of states, this coarsening percolates throughout the other features of their formal system, so that worlds,

[3] Formally, if $\{E_1, E_2, ...E_n\}$ is a set of disjoint events, then $p(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} p(E_i)$.

events and probabilities are also coarsened. Here is how this works. We begin by coarsening the states according to the function $\sigma(\cdot) : S \rightarrow \mathbb{S}$. Next, each coarse-grained world ω is defined as a mapping $T \rightarrow \mathbb{S}$. In other words, a coarse-grained world is a mapping from times to the coarse-grained set of possible states. This yields a set of all nomologically possible coarse-grained worlds $\mathbb{\Omega}$. We then define an algebra $\mathcal{A}_{\mathbb{\Omega}}$ over the set $\mathbb{\Omega}$. A coarse-grained event $\mathbb{E}$ is a set of coarse-grained worlds in $\mathbb{\Omega}$, such that each $\mathbb{E}$ is an element of the algebra $\mathcal{A}_{\mathbb{\Omega}}$. Finally, let $\mathbb{p}_{\boldsymbol{\omega},t}(\cdot)$ be a coarse-grained probability function mapping the coarse-grained algebra $\mathcal{A}_{\mathbb{\Omega}}$ into the unit interval.

To illustrate this account of coarsening, consider a toy example. Suppose that in every nomologically possible world, all that exists is a six-sided die. Once each minute, the die rolls itself, such that it can land showing any of its six sides, including the side that was showing at the previous minute. Recall that each possible world ω is a mapping $T \rightarrow S$, where T is the set of times and S is a set of states. Let each state $s_i \in S$ be a number in the set $S = \{1, 2, 3, 4, 5, 6\}$, where the number corresponds to the side of the die that is facing up. The set of possible mappings of times into this set of states is the set of possible worlds Ω, on which we can define an algebra $\mathcal{A}_{\Omega}$. Let t_1 and t_2 be times in T such that t_1 and t_2 are one minute apart, and let E_{t_2} be the set of possible worlds in which the die is showing six at t_2. If we suppose that, at any given time t_1, the probability that the die will show six at t_2 is $1/6$, then this can be coherently written in List and Pivato's formalism as $p_{\omega,t_1}(E_{t_2}) = 1/6$.

Now suppose that we want to coarsen the set of states, and so we map the fine-grained set of states $S = \{1, 2, 3, 4, 5, 6\}$ into the coarse-grained set of states $\mathbb{S}$={odd, even}. Let us define a coarsening function $S \rightarrow \mathbb{S}$ such that $\sigma(1)$=odd, $\sigma(2)$=even, $\sigma(3)$=odd, $\sigma(4)$=even, $\sigma(5)$=odd, and $\sigma(6)$=even. This yields a coarse-grained set of worlds $\mathbb{\Omega}$ and a coarse-grained event algebra $\mathcal{A}_{\mathbb{\Omega}}$ in which each possible world specifies, for each time, whether the number facing up on the die is odd or even. Let $\mathbb{E}_{t_2}$ be the set of possible worlds in which the die is showing an even number at t_2. If we suppose that, at any given time t_1, the probability that the die will show an even number at t_2 is $1/2$, then this can be written in List and Pivato's formalism as $\mathbb{p}_{\boldsymbol{\omega},t_1}(\mathbb{E}_{t_2}) = 1/2$.

This completes the exposition of List and Pivato's formal framework. The important aspect of the framework to flag here is the way that coarsening is represented. Specifically, it is important to note that List and Pivato coarsen the set of states that the world can be in, and then it is this initial coarsening that implies a coarse-grained set of possible worlds, a coarse-grained algebra of events, and a coarse-grained probability function. In the next section, I will show how this strategy for representing coarsenings plays a key role in List and Pivato's argument for the emergent chance thesis. In subsequent sections, I will advance an alternative formal

account of coarsening that has advantages over List and Pivato's account, and blocks their argument for the emergent chance thesis.

5.3 The Emergent Chance Thesis

Recall that the emergent chance thesis says that it is possible for special science probabilities to be objective chances even if a more fine-grained algebra of events presents an entirely deterministic picture of the dynamics of the natural world. Crucially, this is a claim about what is possibly the case, rather than what is actually the case. As such, proponents of the emergent chance thesis must only establish the possible consistency of coarse-grained objective chances and fine-grained determinism.

The first step in explicating the argument for the emergent chance thesis is to make it more precise. Although List and Pivato follow Schaffer (2007) in offering a wide-ranging discussion of the nature of objective chance, and explicating six desiderata for a concept of objective chance, their core argument for the emergent chance thesis begins with their statement of a sufficient condition for a probability to be *purely epistemic*. This condition is stated as follows.

> **Test for Purely Epistemic Probability**: For a set of possible worlds Ω, let $I_{\omega,t}$ be a set of possible worlds such that every $\omega' \in I_{\omega,t}$ is in the same state as ω up until time t. In world ω and time t, the probability $p_{\omega,t}(E)$ is purely epistemic if $p_{\omega,t}(E|I_{\omega,t}) = 1$ or $p_{\omega,t}(E|I_{\omega,t}) = 0$.[4] (2015, p. 131).

In other words, a non-extreme probability of some event is purely epistemic at a world ω and time t if conditionalizing on the entire history of ω until t renders the probability extreme.

A natural response to this test is to ask why an extreme value of the conditional probability $p_{\omega,t}(E|I_{\omega,t})$ is not *necessary* and sufficient for any non-extreme probability $p_{\omega,t}(E)$ to be purely epistemic. I take it that the answer is as follows. Suppose that we assign unconditional probability $.5$ to a coin coming up heads. Suppose further that if we conditionalize on the full history of the world up until we flip the coin, the conditional probability that the coin comes up heads is $.999$. Under these conditions, there is still some indeterminism to the coin toss, but not much. List and Pivato's sufficient condition for a purely epistemic probability is not satisfied, but one might still want to claim that the initial assignment of $.5$ to the coin coming up heads is purely epistemic. Even though there is some indeterminism with respect to the

[4] The formalism here has been changed slightly, though not consequentially, from the original. This is for the sake of coherence with my exegesis above.

outcome of the coin toss, the initial probability assignment of .5 to the coin coming up heads is solely a representation of ignorance about the coin coming up heads. Thus, a probability $p_{\omega,t}(E)$ can be purely epistemic even when $p_{\omega,t}(E|I_{\omega,t})$ is not extreme. On the other hand, if it is established that $p_{\omega,t}(E)$ is purely epistemic, then we know that it is at least possible that $p_{\omega,t}(E|I_{\omega,t}) = 1$.

Equipped with this sufficient condition for a purely epistemic probability, we can construct an argument for the emergent chance thesis from a formal possibility result:

> **Emergence of Non-Epistemic Probabilities**: It is possible that all non-extreme probabilities in a distribution defined over a fine-grained algebra $\mathcal{A}_\Omega$ are purely epistemic and that some non-extreme probabilities in a distribution defined over the coarse-grained algebra $\mathcal{A}_{\mathbb{\Omega}}$ are not purely epistemic.

This claim is established as follows. Let Ω be a set of possible worlds such that each of world ω is a mapping $T \rightarrow S$. $\mathcal{A}_\Omega$ is the event algebra on Ω. Let it be the case that, for any event $E \in \mathcal{A}_\Omega$, any world ω, and any time t, $p_{\omega,t}(E|I_{\omega,t}) = 1$ or $p_{\omega,t}(E|I_{\omega,t}) = 0$, where the event $I_{\omega,t}$ retains its prior interpretation. Thus, any output of $p_{\omega,t}(\cdot)$ is purely epistemic. All of these assumptions are consistent with a coarsening function $\sigma(\cdot) : S \rightarrow \mathbb{S}$ that yields the coarse-grained set of possible worlds $\mathbb{\Omega}$, the event algebra $\mathcal{A}_{\mathbb{\Omega}}$, and the probability function $\mathbb{p}_{\boldsymbol{\omega},t}$ such that for some coarse-grained events $\mathbb{E}$ and $\mathbb{I}_{\boldsymbol{\omega},t}$, $\mathbb{p}_{\boldsymbol{\omega},t}(\mathbb{E}|\mathbb{I}_{\boldsymbol{\omega},t}) \in (0, 1)$. This establishes the emergence of non-epistemic probabilities.

To illustrate, consider the following example (List and Pivato 2015, p. 136). A set of possible worlds Ω is defined so that each world ω is a function from times to states denoted by real numbers in the interval $[0, 1]$. The history of each world evolves according to a transition rule such that if ω is in state s at t, then ω is in state $f(s)$ at $t + 1$, where $f(s)$ is defined as follows.

$$f(s) = \begin{cases} 2s & \text{if } 0 \leq s \leq 1/2 \\ 2 - 2s & \text{if } 1/2 < s \leq 1 \end{cases} \tag{5.1}$$

In this system, any probabilities defined over the algebra $\mathcal{A}_\Omega$ are purely epistemic. Given some specification of the prior states of any world at any time t, the subsequent evolution of that world is a matter of certainty, such that for any ω, t, and E, $p_{\omega,t}(E|I_{\omega,t}) = 1$ or $p_{\omega,t}(E|I_{\omega,t}) = 0$. For instance, if I_{ω,t_1} is the set of worlds such that $\omega(t_1) = 1/7$, and E is the set of worlds such that $\omega(t_2) = 2/7$, then $p_{\omega,t_1}(E|I_{\omega,t_1}) = 1$.

However, suppose that we map the fine-grained set of states denoted by real numbers in $S = [0, 1]$ into a coarse-grained set of states $\mathbb{S} = \{A, B\}$, via the following function:

$$\sigma(s) = \begin{cases} A & \text{if } 0 \leq s \leq 1/2 \\ B & \text{if } 1/2 < s \leq 1 \end{cases} \tag{5.2}$$

Applying this coarsening function means that if any world is in the fine-grained set of worlds I_{ω,t_1}, i.e. the set of worlds such that $\omega(t_1) = 1/7$, then it is mapped into the coarse-grained set of worlds $\mathbb{I}_{\boldsymbol{\omega},t_1}$, which is the set of worlds such that $\boldsymbol{\omega}(t_1) = A$. Similarly, if any world is in the fine-grained set of worlds E, i.e. the set of worlds such that $\omega(t_2) = 2/7$, then it is mapped into the coarse-grained set of worlds $\mathbb{E}$, which is the set of worlds such that $\boldsymbol{\omega}(t_2) = A$. However, $\mathbb{p}_{\boldsymbol{\omega},t_1}(\mathbb{E}|\mathbb{I}_{\boldsymbol{\omega},t_1}) \neq 1$. To see why, observe that there are many coarse-grained worlds $\boldsymbol{\omega}$ such that $\boldsymbol{\omega}(t_1) = A$ and $\boldsymbol{\omega}(t_2) = B$. Namely, they are just those coarse-grained worlds $\boldsymbol{\omega} \in \mathbb{I}_{\boldsymbol{\omega}t_1}$ that are coarsenings of fine-grained worlds ω in which $\omega(t_1) \in (1/4, 1/2]$. Thus, even though the fine-grained probability distribution $p_{\omega,t}(\cdot)$ is purely epistemic for any ω, t, and E, the coarse-grained probability distribution $\mathbb{p}_{\boldsymbol{\omega},t_1}$ that emerges as the result of applying the coarsening function $\sigma(\cdot)$ to each world in Ω is not purely epistemic for all $\boldsymbol{\omega}$, t, and $\mathbb{E}$.

If we accept the minimal assumption that an objective chance is necessarily represented by a non-extreme, not-purely-epistemic probability, then the emergence of non-epistemic probabilities directly entails the following:

> **Emergent Chance Thesis:** It is possible that all probabilities in a distribution defined over a fine-grained algebra $\mathcal{A}_\Omega$ do not represent objective chances, and that some probabilities in a distribution defined over the coarse-grained algebra $\mathcal{A}_{\mathbb{\Omega}}$ do represent objective chances.

If we let Ω be the set of possible worlds described by some true microphysical theory of the world (i.e. the set of mappings from times to states that specify all physical properties of a world at a given time), and let $\mathbb{\Omega}$ be the set of possible worlds described by some special science (e.g. the set of mappings from times to states that specify all biological properties of a world at a given time), then it is clear how this formal statement of the emergent chance thesis establishes the informal statement of the thesis given in the introduction.

List and Pivato's possibility result depends crucially on their formal way of representing coarse-graining. To see this, consider the following potential response to the case given above. As before, let $\mathbb{E}$ be the set of coarse-grained worlds such that $\boldsymbol{\omega}(t_2) = A$, and let I_{ω,t_1} be the fine-grained set of worlds such that $\omega(t_1) = 1/7$. It

would seem that $\mathbb{p}_{\omega,t_1}(\mathbb{E}|I_{\omega,t_1}) = 1$, i.e. that the probability that a world is such that $\omega(t_2) = A$, given that it is in the fine-grained state $\omega(t_1) = 1/7$, is 1. If this were true, then non-extreme probabilities at both the coarser and finer levels of granularity would both be purely epistemic, and the example would not serve to establish the emergence of non-epistemic probabilities. This counterargument is a formalization of a broader line of response that proceeds as follows. If, according to a fine-grained description of the world, all non-extreme probabilities are purely epistemic, then all non-extreme probabilities assigned to events described at a more coarse-grained level must be purely epistemic. This is because non-extreme probabilities assigned to coarse-grained events reflect ignorance about the fine-grained details of the world.

List and Pivato's response to this line of counterargument is to point out that the probability $\mathbb{p}_{\omega,t_1}(\mathbb{E}|I_{\omega,t_1})$ is not mathematically well-defined. In any conditional probability, the conditioning event and the event being assigned a probability must be in the same algebra of events, which $\mathbb{E}$ and $I_{\omega,t}$ clearly are not (2015, p. 135). This mathematical fact poses a challenge for anyone who aims to counter the emergent chance thesis. Namely, in order to block the emergent chance thesis, we need a formal account of coarsening that satisfies the following three argumentative desiderata: 1) the account must allow for the supervenience and multiple realization relations to hold between events described at finer and coarser levels of granularity, 2) the account must represent fine-grained and coarse-grained events as elements in the same algebra, and 3) the account must have further advantages beyond simply blocking the emergent chance thesis. I clarify that these three conditions are desiderata for an *argument* that blocks the emergent chance thesis, rather than desiderata for any account of higher-level probability. If these conditions were desiderata in the latter sense, then I would be begging the question with respect to List and Pivato's argument, since they reject the second desideratum. In the next section, I present an account of coarsening that satisfies all three of these desiderata.

5.4 An Alternative Account of Coarsening

5.4.1 The Account

The basic formal machinery of my account is unchanged from List and Pivato's. Like them, I define possible worlds as mappings from times to states, and define probability distributions as mappings from the algebra of possible worlds to the unit interval, with each distribution indexed to a particular time and world. Where my account differs from List and Pivato's is in the definition of the coarsening function. Whereas they define their coarsening function $\sigma(\cdot)$ as a mapping from a fine-grained

set of states S to a coarse-grained set of states $\mathbb{S}$, I define a coarsening function $\phi(\cdot) : \mathcal{A}_\Omega \rightarrow \mathcal{A}_\Omega$, i.e. a mapping from the event algebra into itself. Importantly, $\phi(\cdot)$ has to satisfy the constraint that, for any event $E \in \mathcal{A}_\Omega$, it is the case that $E \subseteq \phi(E)$. That is, any event in the algebra must contain only worlds that are also in its coarsening. This constraint is important since it ensures that any fine-grained state is logically consistent with, though not necessarily equivalent to, its own coarsening. Since $\phi(\cdot)$ is a function, supervenience holds; there can be no change at the coarse-grained level without a change at the fine-grained level. Since $\phi(\cdot)$ is possibly not injective, coarse-grained events can be multiply realized by fine-grained events. Thus, my account satisfies the first of the desiderata listed above.

To illustrate how this approach to coarsening works, consider again the simple example of a set of worlds that each contain a single die that repeatedly rolls itself. Each world ω is a mapping from times to a set of states $S = \{1, 2, 3, 4, 5, 6\}$, where the number designating the state corresponds to the side of the die that is facing up. Let E_{ω,t_2} be the event that the die shows a six at t_2, so that $p_{\omega,t_1}(E_{\omega,t_2}) = 1/6$. Now suppose that we want to represent the event that the die shows an even number at t_2. In List and Pivato's framework, this requires us to map each element of the fine-grained set of states $S = \{1, 2, 3, 4, 5, 6\}$ into the coarse-grained set of states $\mathbb{S}$ $= \{\text{odd}, \text{even}\}$. In my framework, we map E_{ω,t_2} into its coarsening $\phi(E_{\omega,t_2})$, where $\phi(E_{\omega,t_2})$ denotes the set of possible worlds in which $s = 2$, $s = 4$, or $s = 6$ at t_2. On this account of coarsening, as in List and Pivato's, we can coherently claim that the probability that the die will show an even number at t_2 is $1/2$, i.e. that $p_{\omega,t_1}(\phi(E_{\omega,t_2})) = 1/2$.

Importantly, my account of coarsening has the same representational power as List and Pivato's. That is, any functional relationship between coarse-grained and fine-grained events that can be represented using their coarsening function $\sigma(\cdot)$ can also be represented by my coarsening function $\phi(\cdot)$. To make this claim more precise, let $\mathcal{A}_\Omega$ be an algebra on the set of possible worlds. Let $\sigma(\cdot)$ be a coarsening function mapping the fine-grained set of states S into the coarse-grained set of states $\mathbb{S}$. If $\mathcal{A}_\mathbb{\Omega}$ is the algebra of coarse-grained events that results from applying the coarsening $\sigma(\cdot)$ to the set of states S, let us say that $\mathcal{A}_\mathbb{\Omega}$ is implied by $\sigma(\cdot)$. Finally, let $\Gamma_{\phi(\cdot)}$ be the range of some coarsening function $\phi(\cdot)$, where $\phi(\cdot)$ is a coarsening function mapping $\mathcal{A}_\Omega$ into itself.[5] The following proposition is true:

Proposition 19. *For any Ω and any $\mathcal{A}_\mathbb{\Omega}$ implied by some coarsening function $\sigma(\cdot)$, there is a coarsening function $\phi(\cdot)$ such that there is a bijection between $\Gamma_{\phi(\cdot)}$ and $\mathcal{A}_\mathbb{\Omega}$.*

[5] That is, $\mathcal{A}_\Omega$ is the codomain of the coarsening function $\phi(\cdot)$, and $\Gamma_{\phi(\cdot)}$ is the range of the function $\phi(\cdot)$.

In other words, for any coarsening produced by List and Pivato's method, we can produce a coarsening via my method that is capable of representing the same coarse-grained events. Thus, any relations of supervenience and multiple realization that are represented by List and Pivato's coarsening function can also be represented via my coarsening function.

5.4.2 Blocking the Emergent Chance Thesis

It should be clear that this account of coarsening also satisfies the second desideratum for a response to List and Pivato, namely that coarse-grained and fine-grained events are represented as elements of the same algebra. Given the extent to which the emergent chance thesis depends on representing coarse-grained and fine-grained events as elements of a different algebra, this formal result suffices to block the argument for the emergent chance thesis. To illustrate, consider the example of emergent chances given above, in which fine-grained states are represented by real numbers in the unit interval, worlds evolve temporally according to the function described by equation (5.1), and the set of states is coarsened according to the function described by equation (5.2). On List and Pivato's framework, the probability that the world is in coarse-grained state A at t_2, given that it is in coarse-grained state A at t_1, is not extreme, and therefore the probability that the world is in coarse-grained state A at t_2 is not purely epistemic. This is meant to hold even though the world's being in coarse-grained state A at t_2 is a logical consequence of the world's being in fine-grained state $2/7$ at t_2, and even though any non-extreme probability assigned to the world's being in fine-grained state $2/7$ at t_2 *is* purely epistemic. Such a result can only hold if we take the fine-grained and coarse-grained events to be elements of a different algebra. As my account represents coarse-grained and fine-grained events as elements of the same algebra, we get the result that if any non-extreme probability assigned to the world's being in fine-grained state $2/7$ at t_2 is purely epistemic, then any probability assigned to the world's being in coarse-grained state A at t_2 is also purely epistemic, a result at odds with the emergence of objective chances in this case.

5.4.3 Further Advantages: The Miners Puzzle

So far, we have two different strategies for representing events at different levels of granularity: List and Pivato's, which entails the emergent chance thesis, and mine, which blocks their argument for the emergent chance thesis. In order to motivate my account, I need to provide some evidence that my account has advantages over List and Pivato's. In what follows, I provide two cases that I believe constitute such evidence. Specifically, I show that my account of coarsening allows us to make better

	Miners in Shaft A	Miners in Shaft B
Block A	10 Saved	0 Saved
Block B	0 Saved	10 Saved
Block Neither	9 Saved	9 Saved

Tab. 5.1: Decision Problem in Miner's Puzzle

sense of two cases that are well known in the philosophy literature: the Miners Puzzle and Simpson's Paradox.

I begin with the Miners Puzzle, which is introduced by Kolodny and MacFarlane (2010). The puzzle imagines a scenario in which ten miners are trapped down one of two mine shafts: shaft A or shaft B. A rainstorm is coming, and rescuers outside the mine have three options. They can block shaft A, block shaft B, or block neither shaft. Blocking one shaft results in the other shaft being completely full of water, while blocking neither shaft results in both shafts being only partially full of water. So, if the miners are in shaft A, then blocking shaft A will save them all, but if they are in shaft B, then blocking shaft A will drown them all. Blocking neither shaft will result in the drowning of one miner, whether the miners are in shaft A or B.

The decision problem facing rescuers can be represented via Table 5.1. Given the structure of the problem, the number of expected lives saved by each action can be calculated by taking the probability that a given action saves any lives, and multiplying it by the number of lives saved. So if the miners are equally likely to be in either shaft, then the expected lives saved by blocking shaft A is $.5(10) = 5$, and the same is true for blocking shaft B. By contrast, the expected lives saved from blocking neither shaft is $1(9) = 9$. So, by consequentialist lights, rescuers should not block either shaft, and thereby save nine lives. Kolodny and MacFarlane go on to argue that on *any* ethical framework, rescuers should block neither shaft.

The Miners Puzzle poses a challenge for accounts of normative decision. It is true that if the miners are in shaft A, then rescuers ought to block A, and the same is true of shaft B. Further, it is stipulated that the miners are either in shaft A or B, and thus it follows that rescuers should either block shaft A or block shaft B. However, it is clear that what they really ought to do is block neither shaft. My aim here is not to address this puzzle directly, but rather to show how List and Pivato's approach to coarsening hamstrings our ability to articulate the subtleties of the puzzle. A crucial fact about the puzzle is as follows. When the rescuers assign probability $.5$ to the event that blocking either shaft will save lives, this probability is purely epistemic. If rescuers learn that the miners are in shaft A, for instance, then the probability

that blocking shaft A saves ten lives goes to 1, and so blocking this shaft becomes more prudent than blocking neither shaft. The same goes for blocking shaft B if the rescuers learn that the miners are in shaft B.

This can be cashed out using List and Pivato's formalism. Let Ω be a set of possible worlds, and let S be the set containing all and only the following states:

- $s_1 =$ 10 miners are alive in shaft A and neither shaft is blocked.
- $s_2 =$ 10 miners are alive in shaft B and neither shaft is blocked.
- $s_3 =$ The rescuers block shaft A and 10 miners are alive in shaft A.
- $s_4 =$ The rescuers block shaft A and 10 miners are drowned in shaft B.
- $s_5 =$ The rescuers block shaft B and 10 miners are alive in shaft B.
- $s_6 =$ The rescuers block shaft B and 10 miners are drowned in shaft A.
- $s_7 =$ The rescuers block neither shaft and nine miners are alive in either shaft.

At t_1, the miners are either in shaft A or B, and the decision about whether to block either shaft is made at t_2. That is, at t_1, any world ω is in state s_1 or s_2, while at t_2, ω is in either s_3, s_4, s_5, s_6 or s_7. Let BA be the set of worlds such that rescuers block shaft A, i.e. worlds such that $\omega(t_2) \in \{s_3, s_4\}$. Let BB be the set of worlds such that rescuers block shaft B, i.e. worlds such that $\omega(t_2) \in \{s_5, s_6\}$, and let BN be the set of worlds such that the rescuers block neither shaft, i.e. worlds such that $\omega(t_2) \in \{s_7\}$. Finally, let X be the set of worlds in which any miners are saved, i.e. worlds such that $\omega(t_2) \in \{s_3, s_5, s_7\}$. For any world ω, $p_{\omega,t_1}(X|BA) = .5$, $p_{\omega,t_1}(X|BB) = .5$, and $p_{\omega,t_1}(X|BN) = 1$. However, to show that $p_{\omega,t_1}(X|BA) = .5$ and $p_{\omega,t_1}(X|BB) = .5$ are purely epistemic, let I_{ω,t_1} be the set of possible worlds that match ω up until t_1. Conditionalizing on this set specifies the state of the world at t_1; that is, it tells rescuers which shaft the miners are in. If the miners are in shaft A, then $p_{\omega,t_1}(X|BA, I_{\omega,t_1}) = 1$ and $p_{\omega,t_1}(X|BB, I_{\omega,t_1}) = 0$; the opposite holds if the miners are in shaft B. Thus, the non-extreme probabilities assigned to the event X, given either BA or BB, are purely epistemic.

This analysis allows us to calculate a response to the following important question: given that rescuers do not know how many miners are in either shaft, how much should they pay to learn this information? The answer is straightforward. In the absence of full information, they will block neither shaft and save nine lives. If they have full information, then they will block either shaft A or shaft B, and save ten

lives. So they should incur costs up to the value of one miner's life in order to learn which shaft the miners are in. Using the formalism developed above, this result can be calculated as follows.

$$10 \cdot p_{\omega,t_1}(X|BA, I_{\omega,t_1}) + 10 \cdot p_{\omega,t_1}(X|BB, I_{\omega,t_1}) - 9 = 1 \tag{5.3}$$

To see the reasoning behind this calculation, note that whatever information is contained in I_{ω,t_1}, rescuers at t_2 will either block shaft A or block shaft B and thereby save ten lives, as compared to the nine lives they would have saved had they blocked neither shaft.

So far, so good for List and Pivato's analysis of the scenario. However, problems arise when we try to adapt their methodology for coarsening to this case. Let $\sigma(\cdot)$ be a coarsening function of the set of states S that is defined as follows.

$$\sigma(s_i) = \begin{cases} s_i & \text{if } i = 1 \\ s_{i-1} & \text{if } i \neq 1 \end{cases} \tag{5.4}$$

In other words, any fine-grained state s_i is mapped to a coarse-grained state s_{i-1}, except for s_1, which is mapped to s_1. Next, we give the following interpretation to all coarse-grained states:

- s_1 = 10 miners are alive in shaft A or shaft B, and neither shaft is blocked.
- s_2 = The rescuers block shaft A and 10 miners are alive in shaft A.
- s_3 = The rescuers block shaft A and 10 miners are drowned in shaft B.
- s_4 = The rescuers block shaft B and 10 miners are alive in shaft B.
- s_5 = The rescuers block shaft B and 10 miners are drowned in shaft A.
- s_6 = Rescuers block neither shaft and 9 miners are alive in either shaft.

Thus, the coarsening function $\sigma(\cdot)$ elides information about which shaft the miners are in at t_1. Let $\mathbb{BA}$ be the event in the resulting coarse-grained algebra such that rescuers block shaft A, i.e. the set of worlds such that $\omega(t_2) \in \{s_2, s_3\}$. Let $\mathbb{BB}$ be the event in the resulting coarse-grained algebra such that rescuers block shaft B, i.e. the set of worlds such that $\omega(t_2) \in \{s_4, s_5\}$. Let $\mathbb{X}$ be the event in the resulting coarse-grained algebra such that any miners are saved i.e. the set of worlds such that $\omega(t_2) \in \{s_2, s_4, s_6\}$. The sufficient condition for a probability to be purely epistemic is not satisfied for $p_{\omega,t_2}(\mathbb{X}|\mathbb{BA}) = .5$ and $p_{\omega,t_2}(\mathbb{X}|\mathbb{BB}) = .5$ within this algebra.

Conditionalizing on the history of the world up to t_2 will not tell rescuers which shaft the miners are in, and therefore will not assign an extreme probability to the event that blocking any shaft saves lives.

One upshot of this analysis is that the question 'what costs should the rescuers be willing to incur to learn what shaft the miners are in?' cannot be well-posed in this coarse-grained algebra. This is because no element of the coarse-grained algebra can specify which shaft the miners are in. Any attempt to do so would require conditionalizing on an element of a more fine-grained algebra to assign an event in a more coarse-grained algebra a conditional probability. In other words, we cannot answer this question because the following equation contains terms with no coherent mathematical definition:

$$10 \cdot \mathbb{P}_{\boldsymbol{\omega},t_2}(\mathbb{X}|\mathbb{BA}, I_{\omega,t_1}) + 10 \cdot \mathbb{P}_{\boldsymbol{\omega},t_2}(\mathbb{X}|\mathbb{BB}, I_{\omega,t_1}) - 9 = 1 \tag{5.5}$$

The mathematical incoherence of these terms is central to List and Pivato's defense of the emergent chance thesis. Thus, List and Pivato are committed to the view that, on the coarsening described in (5.4), the value of the information about which shaft the miners are in cannot be calculated, since the question of which shaft the miners are in is not answered by specifying any element of the coarse-grained algebra. This shows that List and Pivato's account of coarsening, which is central to their argument for the emergent chance thesis, is also liable to produce event algebras that are expressively impoverished in important ways.

By contrast, my account of coarsening is not susceptible to this worry. Recall that the coarsening function $\phi(\cdot)$ is a mapping from the algebra of events $\mathcal{A}_\Omega$ into itself, such that for any $E \in \mathcal{A}_\Omega$, $E \subseteq \phi(E)$. Let the set of worlds Ω be defined as a mapping of times into the seven-element set of fine-grained states specified above. A coarsening function playing the same representational role as the function defined in equation (5.4) can be defined as follows.

$$\begin{aligned}\phi(E) = \{E^* : &\text{ for all } \omega_1 \in E^* \text{ there is an } \omega_2 \in E \text{ such that } \omega_1(t_2) = \omega_2(t_2)) \\ &\text{and there is a pair of worlds } (\omega_3 \in E, \omega_4 \in E^*) \text{ such that } \omega_3(t_1) \neq \omega_4(t_1)\}\end{aligned} \tag{5.6}$$

In other words, $\phi(\cdot)$ maps any set of possible worlds E to a set of possible worlds E^* such that all worlds in E^* agree with some world in E as to the state of the world at t_2, but allows for both possible states of the world at t_1. This ensures that if E is a set of worlds such that the miners are definitively in shaft A, or E is a of worlds such that the miners are definitively in shaft B, then E is mapped to a set of worlds such that the miners are in either shaft A or shaft B. When coarsening is defined in this way, there is no conceptual problem in conditionalizing on the event that the

miners are definitely in shaft A or definitely in shaft B, even though neither event is in the range of the coarsening function $\phi(\cdot)$. This is because these events are still in the algebra over which probability functions are defined. Thus, on my account of coarsening, we can represent the functional relationship between more fine and coarse-grained descriptions of the scenario, while retaining the ability to coherently calculate the maximum cost that rescuers should be willing to take on in order to determine which shaft contains the miners.

This issue with List and Pivato's defense of the emergent chance thesis generalizes beyond the Miner's Puzzle. In many cases, it may be that valuable information is elided when we coarsen the set of possible worlds over which events are defined. From the perspective of these more coarse-grained algebras, it is impossible to coherently say that this information is valuable or that it is missing from our picture of the world. By contrast, my account of coarsening will never run the risk of rendering unintelligible our judgments about the which information is missing from a more coarse-grained picture of the world.

5.4.4 Further Advantages: Simpson's Paradox

Some additional cases that demonstrate the advantages of my account of coarsening are instances of Simpson's Paradox.[6] In general, Simpson's Paradox occurs whenever correlations at one level of description are reversed at a coarser or finer level of description.[7] To illustrate, consider a case in evolutionary biology, which is described in Chuang et al. (2009).[8] Within a population of *Escherichia coli* microbes, individual microbes can either produce or not produce Rhl autoinducer molecules; whether or not a given microbe is a producer of the autoinducer is genetically determined and passed down from a parent microbe to its offspring. These autoinducer molecules improve every microbe in a spatially isolated sub-population's ability to survive and reproduce; one need not be a producer to reap the advantages. However, production of the autoinducer is costly to any given microbe. As such, the best-case scenario for any given microbe is to be a non-producer in a population that is full of producers. The hypothetical evolution of three sub-populations of *Escherichia coli* is described in Table 5.2.[9] It is clear from the table that in each sub-population, non-producers

[6] Similarities between Simpson's Paradox and the Miners Puzzle have been noted by Kotzen (2013).

[7] Although Simpson's Paradox is so-named due its discussion in Simpson (1951), the earliest known discussion of the phenomenon is in Yule (1903). The beginning of philosophical attention to Simpson's Paradox is usually traced to Cartwright (1979). See Malinas and Bigelow (2016) for a summary of Simpson's Paradox and its philosophical significance. It is worth noting that the term 'Simpson's Paradox', though common in the literature, is understood to be a misnomer; strictly speaking, there is no semantic paradox in cases where correlations are reversed at different levels of description.

[8] Similar cases of Simpson's Paradox in evolutionary biology are highlighted by Sober and Wilson (1998).

[9] The expected offspring calculations assume that all microbes die and are replaced by their offspring in the transition from t_1 to t_2.

Population	Phenotype	t_1	t_2	Average Offspring
1	Producer	150	630	4.2
	Non-Producer	50	270	5.2
2	Producer	100	270	2.7
	Non-Producer	100	330	3.3
3	Producer	50	60	1.2
	Non-Producer	150	240	1.6
Total	Producer	300	960	3.2
	Non-Producer	300	840	2.8

Tab. 5.2: Evolutionary Dynamics of *Escherichia coli* Populations

of the autoinducer have an evolutionary advantage over producers. However, in the total population, it appears that producers have an evolutionary advantage over non-producers.

This example shows how in cases of Simpson's Paradox, we see a divergence in the truth values of what Fitelson (2017) calls the "suppositional" and "conjunctive" claims about the relationship between population, autoinducer phenotype, and fitness. To illustrate, consider the following two claims:

> **Suppositional**: If a microbe is in a given sub-population j, then being a non-producer is a more fitness-enhancing strategy than being a producer.

> **Conjunctive**: Being a non-producer and being in sub-population j is a more fitness-enhancing strategy than being a producer in any sub-population.

It should be clear from the data that the suppositional claim is true, but the conjunctive claim is false. In each of the three sub-populations, non-producers of the autoinducer have a greater number of expected offspring than producers; thus, the suppositional claim is true. However, the conjunctive claim is false. For instance, a non-producer microbe in sub-population 3 has lower fitness than a producer microbe in sub-population 1.

If one cannot articulate the distinction between the suppositional and the conjunctive claim, and can instead only evaluate the conjunctive claim, then one loses the ability to make an important distinction about the strategic effectiveness of being a producer or a non-producer. When one sees that the suppositional claim is true while the conjunctive claim is false and assumes that an individual microbe's being a producer

or non-producer does not cause other microbes to be a producer or non-producer, one concludes that a given microbe should adopt the strategy of being a non-producer; conditional on the microbe being in any sub-population, non-production is the best strategy. But if we cannot distinguish microbes by sub-population, then we may be misled into believing the claim that being a producer improves an organism's fitness more than being a non-producer. After all, the apparent truth of the conjunctive claim is supported by the data in the row of Table 5.2 labelled 'Total'. To believe such a claim would be to hold a fundamentally false belief about what would be evolutionarily advantageous to any given microbe in the population. In what follows, I will show that List and Pivato's coarsening method allows us to coarsen the set of possible worlds in a way that would lead us to form just this false belief.

To begin, consider a set of worlds Ω such that each world is a mapping from times t_1 and t_2 to a set of states S such that each state is represented by an ordered triple of ordered pairs $s_i = \langle\langle p_1, np_1\rangle, \langle p_2, np_2\rangle, \langle p_3, np_3\rangle\rangle$ where each p_j and np_j is an integer denoting the number of producers and non-producers in sub-population j. Thus, at time t_1 the population described in Table 5.2 is in state $s_i = \langle\langle 150, 50\rangle, \langle 100, 100\rangle, \langle 50, 150\rangle\rangle$, and at time t_2 the population is in the state $s_i = \langle\langle 630, 270\rangle, \langle 270, 330\rangle, \langle 60, 240\rangle\rangle$. Note that Table 5.2 represents the evolution of one possible world in Ω from t_1 to t_2. Let $F_{np,j>p,k}$ be the set of worlds such that the growth rate for non-producers from t_1 to t_2 in a given sub-population j is greater than the growth rate for producers in a given sub-population k over the same time period (allowing for the possibility that $j = k$). Let $F_{p,j>np,k}$ be the set of worlds such that the growth rate for producers from t_1 to t_2 in a given sub-population j is greater than the growth rate for non-producers in a given sub-population k over the same time period. This allows us to state the following, probabilistic versions of the suppositional and conjunctive claims:

> **Probabilistic Suppositional**: For all worlds ω and sub-populations j, $p_{\omega,t_1}(F_{np,j>p,j}) > p_{\omega,t_1}(F_{p,j>np,j})$.

> **Probabilistic Conjunctive**: For all world ω and all sub-populations j and k, $p_{\omega,t_1}(F_{np,j>p,k}) > p_{\omega,t_1}(F_{p,j>np,k})$.

In other words, the probabilistic suppositional claim states that at t_1, non-producers are more likely than producers to have a higher growth rate from t_1 to t_2 in any sub-population j. If the evolutionary dynamics instantiated in the table above hold in general for Ω, then this claim is true. By contrast, the probabilistic conjunctive claim states that at t_1, it is more likely than not that the growth rate from t_1 to t_2 for non-producers in a sub-population j is greater than the growth rate for producers

from t_1 to t_2 in some other sub-population k. If we take the evolutionary dynamics in the table above to generalize across Ω, then this claim is false.

On List and Pivato's account of coarsening, some coarsenings of this set of possible worlds produce an algebra such that we cannot distinguish between the suppositional and conjunctive claims regarding the evolutionary advantage of being a non-producer. In particular, suppose that we adopt the following coarsening function:

$$\sigma(s_i = \langle\langle p_1, np_1\rangle, \langle p_2, np_2\rangle, \langle p_3, np_3\rangle\rangle) = \mathscr{s}_i = \langle p_1+p_2+p_3, np_1+np_2+np_3\rangle \quad (5.7)$$

In other words, $\sigma(\cdot)$ takes as its input a fine-grained state specifying the number of producers and non-producers in each population, and returns a coarse-grained state specifying the total population of producers and non-producers. As such, this coarsening yields a coarse-grained algebra that does not allow us to distinguish between the growth rates of microbes in any given sub-population. As the comparisons in the growth rates of various sub-populations of microbes are crucial to distinguishing between the suppositional and conjunctive claims, the coarsening function stated in equation (5.7) limits our ability to express crucial aspects of the system under study. Instead, we can only express the claim that in a given microbe population, the growth rate in producers from t_1 to t_2 is likely to be greater than the growth rate in non-producers, a claim that the data in the row of Table 5.2 labelled 'Total' would support. As argued above, this claim is deeply misleading.

As implied by Proposition 19, my account of coarsening can generate a coarsening function with a range that is isomorphic to the set of possible events generated by the coarsening function $\sigma(\cdot)$, as described in equation (5.7). This function can be described as follows:

$$\phi(E) = \{E^* : \text{for all worlds and times in } E \cup E^*, \text{ the sum } p_1 + p_2 + p_3 \text{ takes the same value, and the sum } np_1 + np_2 + np_3 \text{ takes the same value}\} \quad (5.8)$$

That is, $\phi(\cdot)$ maps each event to the set of events that agree with it with respect to the total number of producers and non-producers in all worlds and times. Thus, the coarsening function $\phi(\cdot)$ represents the same supervenience relations between the evolution of the microbe sub-populations and the total microbe population as the coarsening function $\sigma(\cdot)$. However, because $\phi(\cdot)$ is a mapping of the fine-grained algebra of events into itself, we are still able to assign probabilities to events that are distinguished by the composition of particular sub-populations of microbes. As demonstrated above, this is not possible when the coarse-grained algebra of events is produced using the coarsening function $\sigma(\cdot)$. Thus, my account of coarsening does

not threaten our ability to represent important distinctions in cases of Simpson's paradox, whereas List and Pivato's does.

5.5 Response to Counterarguments

An initial line of response to my arguments above could run as follows. In considering both the Miners Puzzle and Simpson's Paradox, I have defined List and Pivato-style coarsening functions that undercut our ability to articulate important aspects of either case, and taken this as evidence for the deficiency of their approach. However, rather than taking these examples to be bad results for List and Pivato's account of coarsening, why not simply take them to be bad results for the particular choice of coarsening function? The success of the special sciences, this argument continues, is due at least in part to the fact that those sciences coarsen the physical description of the world in the right way. Thus, in contrast with the examples that I have described above, in well-formulated special sciences we are able to express all important nuances of all relevant cases. According to this argument, the coarsening functions used in these special sciences will be just those that are able to express the important details of all domain-specific cases, while still using a coarse-grained event algebra.

In response, I argue that List and Pivato's coarsening method lacks the resources to articulate whether or not a given coarsening is deficient in the ways described above. Consider the case of the Miner's Puzzle. If we coarsen the set of possible worlds so as to elide information about which shaft the miners are in, then we lose the ability to say that this is information that we would like to have, if we could get it. One could tell a similar story with respect to the example of Simpson's Paradox that I have given above. By coarsening the set of possible worlds so as to elide information about the sub-population that the organism is in, we end up thinking that it is better for a microbe to produce the autoinducer than to not produce it. Further, from within this coarsening, there is no way to say that the coarsening is in any way deficient, or that sub-population-level statistics on the fitness of microbes would be valuable to a microbe that could understand them. In summary, absent some meta-algebra representing different possible coarsenings, List and Pivato's formalism lacks any way of explicitly comparing coarsenings. Further, even if such a meta-algebra were to be supplied, it is unclear what would be gained from such an algebra that one does not already get from the more parsimonious option of representing coarsening as a mapping from the event algebra into itself.

A second line of objection to what I have presented here stresses the fundamental motivation behind the emergent chance thesis. This argument proceeds as follows. At the level of description used in any given special science, events may be assigned

non-extreme probabilities, and those probabilities may not be explained by any agent's lack of information about the world *as described by that special science*. These probabilities, the objection goes, deserve the name "objective chances" within the context of a given special science. List and Pivato's formal framework allows this to be articulated, whereas my account seems to imply that the only non-extreme probabilities that qualify as objective chances are those events that have non-extreme probability, conditional on a maximally fine-grained description of the world up to some relevant time.

In response, I would argue that my approach does leave open the possibility of a kind of special science objective chance, albeit one that is conceived of in a different way than List and Pivato have in mind. To illustrate, consider the example of a wheel of fortune that can stop on white or red. Suppose that the color that the wheel stops on is completely determined by the amount of force with which the wheel is spun. For instance, if the wheel is spun with anywhere from 14.0 to 14.1 Newtons of force, then it will land red, if it is spun with anywhere from 14.1 to 14.2 Newtons of force, then it will land white, and so on. If the wheel is spun at t_1, and I_{ω,t_1} specifies the force with which the wheel is spun, then $p_{\omega,t_1}(R|I_{\omega,t_1}) = 1$ or 0 for all ω. As such, if the probability assigned to the event that the wheel stops on red with any given spin is non-extreme, then it is purely epistemic. However, suppose that, as a matter of fact, actual spins of the wheel vary in force such that the wheel stops on red about half the time. Let $\phi(I_{\omega,t_1})$ be the event that the wheel is spun, with any force, at time t_1. Under these conditions, $p_{\omega,t_1}(R|\phi(I_{\omega,t_1})) = .5$. In the sense articulated by List and Pivato, this probability is purely epistemic, since it reflects our ignorance about the amount of force applied in any given spin. However, one could argue that the probability is still objective in the sense that it reflects a mind-independent fact about the distribution of possible amounts of force exerted in any given spin of the wheel. In the same way, if events in the special sciences are realized by microphysical conditions in a way that gives rise to non-extreme probabilities at a coarse-grained level of description, then perhaps they can also be thought of as objective in some sense, even if they are purely epistemic by List and Pivato's lights. A fully-fledged defense of this proposal would require substantially more argumentation, but I have outlined the proposal here so as to highlight the fact that rejecting List and Pivato's approach does not necessarily undermine our ability to treat special-science probabilities as objective in some sense. For proposals for understanding special-science probabilities that point in this direction, see Shalizi and Moore (2003), Lyon (2011), and Strevens (2011).

5.6 Connection to Previous Chapters

This chapter has followed List and Pivato in using a different formalism than was used in the previous chapters. It also does not consider Bayesian networks. However, the conclusions defended here are consonant with the formal approach to coarse-graining taken in the previous chapters. The central formal upshot of my argument here can be understood as follows. Whereas List and Pivato model coarsening as a change in the outcome space Ω, an algebra of which probabilities are defined on, I model coarsening while preserving the same underlying outcome space. I do this by defining a surjective but not necessarily injective function from the algebra $\mathcal{A}_\Omega$ into itself, so that only larger subsets of Ω are salient.

List and Pivato do not explicitly frame their proposal in terms of probability spaces and random variables, but their arguments imply a formalization that proceeds as follows. First, we start with a fine-grained probability space $\langle \Omega, \mathcal{A}_\Omega, p_{\omega,t}(\cdot)\rangle$. This allows us to assign probabilities to any random variable $X : \Omega \rightarrow R_X$. Now, on List and Pivato's framework, if we wish to coarsen our model of the world, we switch to a new probability space $\langle \mathbb{\Omega}, \mathcal{A}_{\mathbb{\Omega}}, \mathbb{p}_{\omega,t}(\cdot)\rangle$. This yields new, coarse-grained random variables $\mathbb{X} : \mathbb{\Omega} \rightarrow \mathbb{R}_{\mathbb{X}}$. By contrast, I begin with the same probability space $\langle \Omega, \mathcal{A}_\Omega, p_{\omega,t}(\cdot)\rangle$ and random variables of the form $X : \Omega \rightarrow R_X$, but in coarsening my representation of the world, I simply coarsen the value set of the random variables used in models. One can understand my coarsening function $\phi(\cdot) : \mathcal{A}_\Omega \rightarrow \mathcal{A}_\Omega$ as a coarsening function on the cross-product of the value sets of all random variables used in a model. The function $\phi(\cdot)$ specifies how events are represented at a more coarse-grained level, without changing the underlying outcome space or algebra over which probabilities are defined. This is all in keeping with my stipulation, in the introduction to this thesis, that all random variables in any Bayes net are defined over the same outcome space of possible worlds. As proven in this chapter, this more formally parsimonious framework for coarse-graining has just as much representational power as List and Pivato's.

5.7 Conclusion

Whether the sorts of special science probabilities described in the previous paragraph are genuine emergent chances strikes me as a largely verbal dispute about what counts as an objective chance. If an objective chance is a non-extreme probability that does not reflect any ignorance about the state of the world up until some time, then I would claim, in contrast with List and Pivato, that there are only objective chances if the underlying microphysics is indeterministic. However, if objective chances are just those probabilities that play an important role in the special sciences, then

clearly such objective chances exist, regardless of our preferred formalization of the relationship between coarse-grained and fine-grained events.

Importantly, even if it turns out that special science probabilities are purely epistemic, it does not follow from this that they are not interesting in their own right. In studying the inter-theoretic relations between various sciences, we discover which fine-grained facts about a given target system can be systematically ignored, while still retaining all domain-relevant information. For instance, in studying the relationship between thermodynamics and statistical mechanics, we discover that if we only care about the thermodynamic properties of that gas, then specific information about the velocities of each particle in a box of gas can be ignored. These facts about which pieces of information get included in more coarse-grained descriptions of the world are highly informative as to the nature of the subject matter studied by these more coarse-grained theories. This is true independently of whether the probabilities included in those special science theories are objective chances, and independently of any particular formal representation of coarse-graining.

5.8 Appendix

5.8.1 Proof of Proposition 19

Proof. Let $\omega \in \Omega$ be any possible world. Recall that ω is a mapping $T \rightarrow S$. Let a "conjoined pair" of states $(s_i, s_j) \in S$ be any pair of states such that, according to the coarsening function $\sigma(\cdot)$, $\sigma(s_i) = \sigma(s_j) = \mathbb{s}$. Let $\Omega'_k \subseteq \Omega$ be a set of possible worlds such that, for any pair of worlds $(\omega_l, \omega_m) \in \Omega'_k$ and all t, the states $\omega_l(t)$ and $\omega_m(t)$ are conjoined. The set of such sets of worlds $\Omega^\dagger = \{\Omega'_1, \Omega'_2, ..., \Omega'_n\}$ fully partitions the set of all possible worlds Ω. If the coarsening function $\sigma(\cdot)$ is applied to the set of states S, then each set of fine-grained worlds Ω'_k is replaced by a single coarse-grained world $\boldsymbol{\omega}$. Thus, we can define a bijection from $\Omega^\dagger$ into the set of coarse-grained possible worlds $\mathbb{\Omega}$, and therefore a bijection from $\mathcal{A}_{\Omega^\dagger}$ into $\mathcal{A}_{\mathbb{\Omega}}$.

Finally, define a coarsening function $\phi(\cdot) : \mathcal{A}_\Omega \rightarrow \mathcal{A}_\Omega$ such that $\mathcal{A}_{\Omega^\dagger}$ is the range of $\phi(\cdot)$, i.e. $\Gamma_{\phi(\cdot)} = \mathcal{A}_{\Omega^\dagger}$. For any $E \in \mathcal{A}_\Omega$ and any $E^\dagger \in \mathcal{A}_{\Omega^\dagger}$, if $\omega \in E$, then $\omega \in E^\dagger$. Thus, in keeping with my proposed restriction on any coarsening function $\phi(\cdot)$, for any $E \in \mathcal{A}_\Omega$, $E \subseteq \phi(E)$. It follows that there is a bijection between $\Gamma_{\phi(\cdot)}$ and $\mathcal{A}_{\mathbb{\Omega}}$. □

6 Imprecise Bayesian Networks as Causal Models

6.1 Introduction

Throughout this dissertation, I have been concerned with cases in which scientists represent systems in nature using causal models with coarse-grained random variables. That is, I have been concerned with why and how scientists choose coarser representations of the types of events that can occur within a given system. However, this is not the only dimension according to which a causal model can be coarsened. We might also coarsen the joint probability distribution over the cross product of the value set of each variable in a Bayes net. Instead of using a single joint probability distribution, we might represent the probabilistic relationships between variables in a model using a *set* of joint probability distributions. Such a representation is appropriate whenever the uncertainty of our beliefs about a system or the stochasticity of the system itself is severe enough that it cannot be represented by a single probability distribution. This attitude of severe uncertainty with regard to the possible states of a system being modeled is often called *ambiguity*.

Probabilistic models in many scientific contexts often represent their target systems in a way that is characterized by ambiguity. Consider the following real-world example, due to Antonucci et al. (2004). Geologists in the Ticino canton in southern Switzerland want to know how likely it is that a debris flow in a given region will be of low, medium, or high severity, given certain information about the geomorphology of a region. A debris flow is a geological incident that is similar to a mudslide. A low-severity flow is one in which the thickness of the debris displaced is less than 10cm, a medium-severity flow is one in which the thickness of the debris displaced is between 10cm and 50cm, and a high-severity flow is one in which the thickness of the debris displaced is greater than 50cm. A statistical model representing the relationship between the geomorphology of a region and the severity of a debris flow contains several input variables, such as land use and antecedent soil moisture. Scientists are uncertain as to the values of these variables, and they are also uncertain about the extent to which these values influence the probability distribution over the severity of a debris flow. The model reflects this uncertainty by assigning probability intervals, rather than precise probabilities, to each of the possible debris flow outcomes.

To illustrate, for one set of inputs, Antonucci et al.'s model determines that the probability of a low-severity flow is in the interval $[.08, .30]$, that the probability of a medium-severity flow is in the interval $[.23, .32]$, and that the probability of a high severity flow is in the interval $[.46, .69]$ (Antonucci et al. 2004, p. 7).[1]

It is plausible that the inherent ambiguity of these probabilistic outputs is a feature, not a bug, of the model that generates them. As Joyce (2010) argues, there are cases in which agents ought to have imprecise degrees of belief with respect to future outcomes, since their total evidence does not warrant the assignment of a precise probability to any particular outcome.[2] I take it that the geological case described above falls into this category. Indeed, Antonucci et al. claim that the use of probability intervals, as opposed to precise probabilities, allows them to "quantify uncertainty on the basis of historical data, expert knowledge, and physical theories" (2004, p. 1). Further, the uncertainty regarding the relationships between various inputs and the severity of a debris flow is such that one cannot define a second-order probability distribution over the interval range. Thus, it does not make sense to take an average over the interval and arrive at a single, precise probability. Rather, the imprecise judgments produced by the model may well represent the epistemic attitude that is most warranted by the available evidence.

As I have emphasized throughout this dissertation, Bayesian network models such as those developed by Pearl (2000) and Spirtes et al. (2000) provide powerful tools for representing the causal structure of the world. Traditionally, these models exclusively use precise probabilities. Thus, it is an important question whether extensions of these models that allow for imprecise probabilities will yield similar fruit. Most work on imprecise Bayesian networks (IBNs) has occurred in the statistics and computer science literature; see Corani et al. (2012) for a comprehensive overview.[3] Unsurprisingly, this literature has focused largely on the use of IBNs to make predictions about future events. By contrast, there has been little attention focused on the philosophical issue of whether and to what extent IBNs can be used to represent the causal structure of actual systems; that is, whether the edges in an IBN can be interpreted as type-level causal relations between variables.[4] Addressing this issue is my focus in this chapter.

My central task is to determine whether and how the Causal Markov Condition (CMC) condition can be extended into the imprecise context. In general, determining

[1] In the example cited, these probability intervals are calculated using the imprecise Bayes nets techniques discussed in this chapter.

[2] See Bradley (2016) for a full accounting of the various reasons discussed in the literature for adopting epistemic attitudes best represented by imprecise probabilities.

[3] In the computer science and statistical literature, IBNs are sometimes called *credal networks*. I avoid this term here because of the different usage of the term 'credence' in formal epistemology.

[4] The meaning of 'edge' and 'variable' in this context will be made precise in the next section.

whether or not a Bayes net satisfies CMC requires an understanding of what it means for two random variables to be independent of one another. In the precise context, such independence has only one mathematical definition, but in an imprecise context, multiple independence concepts abound (see Couso et al. 1999, Cozman 2012). Most saliently, imprecise probability distributions over different variables may satisfy so-called *complete independence*, or the weaker notion of *epistemic independence*. In what follows, I argue that neither complete nor epistemic independence can be neatly "plugged in" to CMC to yield a compelling set of adequacy conditions for the causal interpretation of IBNs without troubling implications. These implications can be avoided by positing further restrictions on the probabilistic features of IBNs, but such restrictions demonstrate the extent to which many otherwise innocuous imprecise probabilistic models cannot be interpreted causally. Thus, I conclude that introducing imprecision into probabilistic graphical models can limit their ability to represent causal structure.

My plan for this chapter is as follows. In Section 6.2, I present the basic formalism for precise Bayesian network models, show how CMC and Faithfulness license a causal interpretation of those models, and describe the central role that probabilistic independence plays in defining these conditions. In Section 6.3, I present the basic formalism for IBNs. In Section 6.4, I illustrate the distinction between complete and epistemic independence between random variables in an imprecise probabilistic context. In Section 6.5, I argue that neither complete nor epistemic independence can be used to adapt CMC and Faithfulness into an imprecise context. In Section 6.6, I consider and respond to some salient objections to my argument. In Section 6.7, I offer concluding remarks.

6.2 Precise Bayes Nets as Causal Models

Recall from Chapter 1 that a Bayes net is a triple $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ where $\mathcal{V}$ is a set of random variables, $\mathcal{E}$ is a set of directed edges relating the variables in $\mathcal{V}$, and $p(\cdot)$ is a joint probability distribution over the cross product of the value set of each variable in $\mathcal{V}$. Recall also from Chapter 1 that since a Bayes net satisfies CMC, if it also satisfies Minimality, then: 1) all variables are probabilistically independent of their more distant ancestors, given their parents, 2) all correlated variables are related via a path of directed edges or have a common ancestor, and 3) the Bayes net contains no excess directed edges. Thus, if a Bayes net satisfies CMC (as all Bayes nets do by definition) and Minimality, then we can interpret its directed edges causally.

However, note that in Chapter 1, I still had to stipulate that a given Bayes net was an adequate representation of the causal structure of its target system. This is because, even though satisfying CMC and Minimality licenses a causal interpretation of a Bayes net, this does not entail that any Bayes net that satisfies these conditions provides a *correct* representation of a particular system. To see this, suppose that the variables X, Y, and Z are all correlated with one another. The graphs $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$, and $X \leftarrow Y \rightarrow Z$, among others, will all satisfy CMC and Minimality. In order to determine which graph is correct, we will need to perform experiments; for example, if we exogenously fix the value of the variable X and this changes the probability distribution over Z, then $X \rightarrow Y \rightarrow Z$ is the correct graph of the three listed above. While the discovery of the correct causal graph using interventions is an important part of the literature on causal Bayes nets, I will bracket this discussion for the remainder of this chapter. What is important for my purposes is that, because they satisfy CMC and Minimality, all three of these graphs can be interpreted as hypotheses about the causal structure of the world.

However, assuming that the correct Bayes net representation of some data set satisfies CMC and the stronger Faithfulness condition can yield some definitive conclusions regarding the causal structure of target systems. Let (X, Y) be a pair of adjacent variables in a Bayes net $\mathcal{N}$ if and only if $\mathcal{N}$ contains an edge $X \rightarrow Y$ or an edge $Y \rightarrow X$. Meek (1995) extends a result from Verma and Pearl (1990a) proving that if two Bayes nets share the same variable set $\mathcal{V}$ and the same joint probability distribution $p(\cdot)$, where $p(\cdot)$ is *Faithful* to the graph in questions, then they share the same pairs of adjacent variables. Faithfulness in this context is defined as follows:

> **Faithfulness:** A joint probability distribution $p(\cdot)$ is faithful to a graph $\mathcal{G}$ if and only if all variables in $\mathcal{G}$ are independent of only their non-descendants, given their parents.

Note that Faithfulness is strictly stronger than Minimality. Moving forward, I will assume that any Bayes net must satisfy the stronger Faithfulness assumption in order to represent the causal structure of its target system. Thus, for any given precise probability distribution $p(\cdot)$ and variable set $\mathcal{V}$, all Bayes nets that are consistent with $p(\cdot)$ will agree with respect to which variables are directly causally related and which variables are not. The ambiguity as to the correct Bayes net, given the joint probability distribution over $\mathcal{V}$, is solely a matter of the *direction* of the direct causal relationship between two variables; there is no ambiguity as to whether such a relationship exists. To put this another way, if it were to turn out to be the case that no direct causal connection in fact exists between two variables related by a direct edge in a Bayes net, it would have to be the case that either of two conditions obtains: 1) there is a failure of causal sufficiency, i.e. there are latent variables that

have not been included in the Bayes net,[5] 2) the probability distribution over the existing variables is inaccurate, or 3) the probability distribution violates Faithfulness. Similarly, if there *is* a direct causal connection between two variables that are not related by a directed edge, then there is a either a failure of causal sufficiency or the joint probability distribution over the graph is inaccurate. This important feature of Bayes nets turns out not to hold in the imprecise context, and this contrast between precise and imprecise Bayes nets will be important to my subsequent analysis.

This analysis helps itself to the existence of a model-independent fact of the matter as to whether there is a causal relationship between two variables. Though the nature of such model-independent causal relationships is a topic well beyond the scope of this chapter, for the sake of argument I adopt a mechanistic understanding of such relationships; one variable directly causes another if there is a direct mechanistic connection between changes in the value of one variable and changes in the value of the other. Following Machamer et al. (2000), I understand mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (2000, p. 3). As such, a mechanistic connection between two variables means that, in the actual world, there are entities and activities organized such that changing the value of one variable leads to regular changes in the probability distribution over another variable. A direct mechanistic connection is one that holds independently of the values taken by the other variables in the graph. This is a decidedly counterfactual or interventionist understanding of the meaning of 'productive', in keeping with the understanding of mechanisms advanced by Woodward (2002). This stands in contrast to a more ontic notion of production advocated by Glennan (2002), according to which production in mechanisms requires a physical process connecting the components of a mechanism.

Before moving to discuss Bayes nets with imprecise probabilities, I must first address the concept of probabilistic independence between random variables. Obviously, probabilistic independence plays a crucial role in our understanding of CMC and Faithfulness, since independence is part of the definition of both . Therefore, one must understand the relation of independence between random variables in order to understand the causal interpretation of a Bayes net. Where all probabilities are assumed to be precise, independence has a straightforward mathematical meaning, which can be stated as follows.

> **Precise Unconditional Independence:** X and Y are independent if and only if, for all values x_i and y_j, $p(x_i, y_j) = p(x_i) \cdot p(y_j)$.

[5] A precise definition of causal sufficiency is as follows. For any variable X that is not in $\mathcal{V}$ and is a common cause of two or more variables in $\mathcal{V}$, all variables in all $\mathcal{V}$ are independent of X (Spirtes et al. 2000, p. 475).

Similarly, conditional independence in the precise probabilistic context can be defined as follows.

> **Precise Conditional Independence:** X and Y are independent conditional on Z if and only if, for all values x_i, y_j and z_k, $p(x_i, y_j|z_k) = p(x_i|z_k) \cdot p(y_j|z_k)$.

Different modellers may hold that different statistical hurdles have to be cleared to license the conclusion that there is dependence or independence between two variables. For example, different modellers may hold that data must support the existence of a dependence relation between two variables at a certain level of statistical significance. Nevertheless, where precise probabilities are used, the *mathematical* definitions of independence and conditional independence are just those given above.

6.3 Imprecise Bayes Nets: The Basics

Following Corani et al. (2012), I define an imprecise Bayes net as a triple $\mathcal{I} = \langle \mathcal{V}, \mathcal{E}, \mathcal{K} \rangle$. As in a precise Bayes net, $\mathcal{V}$ is a set of variables and $\mathcal{E}$ is a acyclic set of directed edges. However, whereas a precise Bayes net contains a joint probability distribution $p(\cdot)$, an imprecise Bayes net contains a set of joint probability distributions $\mathcal{K}$. We can use $\mathcal{K}$ to derive sets of conditional and marginal probabilities over the variables in the graph that are consistent with the set of joint distributions $\mathcal{K}$. Following Levi (1974, 1980), I call these sets of probability distributions over variables "credal sets".[6] For example, if X is a variable in an IBN, then the set of marginal probability distributions over X that are consistent with the set of joint distributions $\mathcal{K}$ is the credal set for X. If Y is another variable in the graph, then the set of conditional probability distributions over X, given that $Y = y_j$, that are consistent with $\mathcal{K}$ is the conditional credal set over X, given that $Y = y_j$.

To illustrate via a toy example, suppose that an IBN contains a binary variable S denoting whether a person smokes, and another binary variable LC denoting whether they develop lung cancer. If the probability that someone develops lung cancer, given that they smoke, is best represented by the interval $[.05, .2]$, then we can define the following conditional credal set over LC, given $S = 1$ (i.e. given that the person smokes):

$$\mathcal{K}_{S=1} = \{p(\cdot) : p(LC = 1|S = 1) \in [.05, .2],\ p(LC = 0|S = 1) \in [.8, .95]\} \quad (6.1)$$

[6] Importantly, I do not endorse, nor do I deny, the subjective interpretation of these sets that philosophical readers will associate with the term 'credal'.

This is the set of all precise probability distributions over LC, given $S = 1$, such that any probability $p(LC = 1|S = 1)$ that is consistent with $\mathcal{K}_{S=1}$ takes a precise value within the interval $[.05, .2]$, and such that any probability $p(LC = 0|S = 1)$ that is consistent with $\mathcal{K}_{S=1}$ takes a precise value within the interval $[.8, .95]$. From this element-by-element, or *pointwise*, analysis of the credal set, we can infer the interval-valued claim that the probability of lung cancer, given smoking, is in the interval $[.05, .2]$.

This method of representing imprecise probabilities as sets of probability distributions has its roots in foundational work by Levi (1974, 1980) and Walley (1991). However, unlike these authors, I do not assume that all credal sets are convex. That is, I drop the assumption that for any two probabilities in the credal set, the linear average of these two probabilities is also in the credal set. While perhaps non-standard, the decision to eschew the convexity requirement for credal sets has some precedent in the recent literature; see for instance Seidenfeld et al. (2010) and Elkin and Wheeler (2018).

One further explanation of notation is required. Throughout this chapter, I refer to the set of marginal probability assignments for some variable value x_i that are consistent with the probability distributions in some credal set $\mathcal{K}$ using the notation $\mathcal{K}[x_i]$. More precisely, $\mathcal{K}[x_i]$ can be defined as follows:

$$\mathcal{K}[x_i] = \{n : \text{there exists a } p(\cdot) \in \mathcal{K} \text{ such that } p(x_i) = n\} \tag{6.2}$$

Similarly, I refer to the set of conditional probability assignments for some variable value x_i, given some other variable value z_k, that are consistent with the probability distributions in some credal set $\mathcal{K}$ using the notation $\mathcal{K}[x_i|z_k]$. More precisely, $\mathcal{K}[x_i|z_k]$ can be defined as follows:

$$\mathcal{K}[x_i|z_k] = \{n : \text{there exists a } p(\cdot) \in \mathcal{K} \text{ such that } p(x_i|z_k) = n\} \tag{6.3}$$

This notation will be used frequently throughout the remainder of this chapter.

To interpret an IBN causally, we will need to extend the notions of unconditional and conditional independence between variables to apply to cases where there is not a precise joint probability distribution over all variables in a graph, but where there is a specified family of joint probability distributions over the variables in a graph. This would allow us to define imprecise extensions of the CMC and Faithfulness conditions. However, as described in the introduction, there is no single notion of what it means for two variables to be independent when independence is defined in terms of imprecise probabilities. In the next section, I will outline the two leading ways of cashing out independence in an imprecise context: complete independence

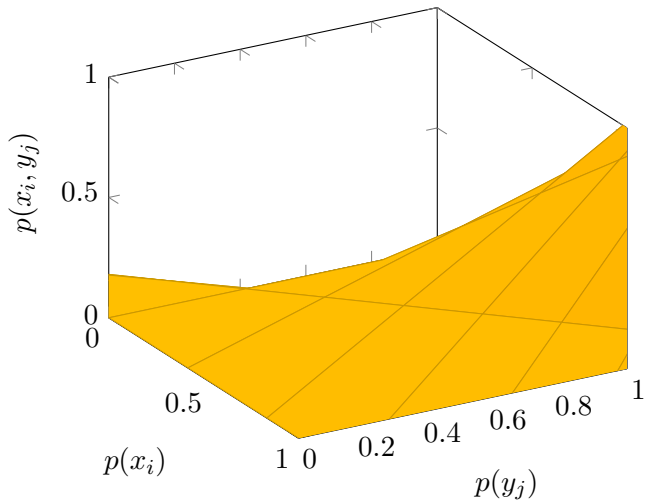

Fig. 6.1: The Set $\{p(\cdot) : p(x_i, y_j) = p(x_i) \cdot p(y_j)\}$ for values $X = x_i$ and $Y = y_j$

and epistemic independence. I will show that there are substantive differences between the two notions of independence, with important consequences for the causal interpretation of IBNs.

6.4 Two Concepts of Independence

6.4.1 Complete Independence

The most natural extension of the concept of probabilistic independence into the imprecise context is via the notion of what Cozman (2012) calls "complete independence". A simple statement of complete independence is as follows.

> **Complete Independence:** X and Y are completely independent according to credal set $\mathcal{K}$ if and only if $\mathcal{K} \subseteq \{p(\cdot) : p(x_i, y_j) = p(x_i) \cdot p(y_j)$ for all x_i and $y_j\}$.

To illustrate, consider the graph in Figure 6.1, on which the z-axis values represent the joint probability $p(x_i, y_j)$ and the x and y axis values represent the marginal probabilities $p(x_i)$ and $p(y_j)$, respectively. The surface shown in the graph is the set of joint probabilities such that $p(x_i, y_j) = p(x_i) \cdot p(y_j)$. If the set of joint probabilities $\mathcal{K}[x_i, y_j]$ is a subset of this set for all values x_i and y_j, then X and Y are completely independent.

It is worth noting that many discussions of independence in the imprecise context would start with the notion of "strong independence"; this independence concept "has been adopted by most of the authors who have modeled independence using

imprecise probabilities" (Couso et al. 1999, p. 7). Two variables X and Y are strongly independent if and only if for all values x_i and y_j, the set of joint probability distributions $p(x_i, y_j)$ are in the *convex hull* of the set $\{p(\cdot) : p(x_i, y_j) = p(x_i) \cdot p(y_j)\}$, i.e. the smallest convex set containing the set $\{p(\cdot) : p(x_i, y_j) = p(x_i) \cdot p(y_j)\}$. This move has its roots in Levi (1980), and is required in order to maintain the convexity of credal sets, as the set $\{p(\cdot) : p(x_i, y_j) = p(x_i) \cdot p(y_j)\}$ is not generally convex. However, since I do not require that credal sets be convex, strong independence is less relevant to my discussion here. One could substitute complete independence for strong independence and run roughly the same argument that I make in this chapter, but this would ultimately be more mathematically interesting than philosophically interesting.

This definitions can be straightforwardly augmented to provide a definitions of complete conditional independence.

> **Conditional Complete Independence:** X and Y are completely independent conditional on Z according to a credal set $\mathcal{K}$ if and only if $\mathcal{K} \subseteq \{p(\cdot) : p(x_i, y_j|z_k) = p(x_i|z_k) \cdot p(y_j|z_k)$ for all x_i, y_j and $z_k\}$.

It should be clear from the previous section that this notion of conditional complete independence will be crucial for understanding how the CMC and Faithfulness conditions can be extended into the imprecise context, if complete independence is taken to be the operative independence concept for a causal interpretation of an imprecise Bayes net.

Complete independence can also be understood as a pointwise notion of independence. Recall that the joint credal set over X and Y is a set of precise joint probability distributions. Complete independence requires that every element (or "point") in this set satisfy the precise version of independence or conditional independence. As I will demonstrate shortly, this same pointwise analysis is not applicable to all notions of independence in the imprecise context.

6.4.2 Epistemic Independence

A rival notion of independence between two random variables in the imprecise context is the notion of epistemic independence; according to Walley (1991), this is the correct way of understanding independence between variables in an imprecise context. To define epistemic independence, recall that $\mathcal{K}[x_i]$ is a set of marginal probabilities assigned to $X = x_i$ by the distributions in $\mathcal{K}$, and that $\mathcal{K}[y_j]$ is the set of marginal probabilities assigned to $Y = y_j$ by the distributions in $\mathcal{K}$. We can define epistemic independence as follows.

Epistemic Independence: X and Y are epistemically independent according to some credal set $\mathcal{K}$ if and only if $\mathcal{K} \subseteq \{p(\cdot) : p(x_i|y_j) \in \mathcal{K}[x_i],\ p(y_j|x_i) \in \mathcal{K}[y_j]$ for all x_i and $y_j\}$.

This definitions can be straightforwardly adapted to the conditional context. Recall that $\mathcal{K}[x_i|z_k]$ be a set of conditional probabilities assigned to $X = x_i$ by the distributions in $\mathcal{K}$ when $Z = z_k$, and $\mathcal{K}[y_j|z_k]$ be a set of conditional probabilities assigned to $Y = y_k$ by the distributions in $\mathcal{K}$ when $Z = z_k$. Conditional epistemic independence can then be defined as follows:

Conditional Epistemic Independence: X and Y are epistemically independent conditional on Z according to some credal set $\mathcal{K}$ if and only if $\mathcal{K} \subseteq \{p(\cdot) : p(x_i|y_j, z_k) \in \mathcal{K}[x_i|z_k],\ p(y_j|x_i, z_k) \in \mathcal{K}[y_j|z_k]$ for all x_i, y_j and $z_k\}$.

These definitions lay the groundwork for a possible causal interpretation of IBNs in which epistemic independence is the operative independence concept.

Determining whether a given joint probability distribution is within the set of joint distributions that satisfy epistemic independence is computationally difficult, requiring multi-linear programming techniques (see Campos and Cozman (2007)). Further, unlike complete independence, epistemic independence does not impose a particular independence condition on every element of the joint credal set. That is, it is not a pointwise notion of independence. Rather, it is a *setwise* notion of independence, requiring only that the whole joint credal set $\mathcal{K}$ is such that all of the conditional probability distributions over each variable that are consistent with $\mathcal{K}$ are within all of the marginal probability distributions over each variable that are consistent with $\mathcal{K}$.

6.4.3 Distinguishing the Two Independence Concepts

Complete independence is a strictly stronger condition on the probabilistic relationship between two variables than epistemic independence (see Couso et al. (1999) for a demonstration of this point). This means that complete independence implies epistemic independence, but not vice versa. However, more interesting distinctions between the two concepts can also be drawn. Namely, satisfying complete independence seems to require the lack of a mechanistic connection relation between the values of two variables. By contrast, satisfying epistemic independence requires a purely informational or evidential independence between the two variables. That is, if two variables are epistemically independent, then learning the value of either variable should not change the behavior of an agent, where that behavior depends on the value of the other variable (e.g. deciding how much to gamble on the value of

either variable). To illustrate this distinction, consider the following toy case, which is adapted from Couso et al. (1999):[7]

> **Urn Example:** There are three urns, A, B, and C, each of which contain 100 balls. The balls in each urn are either red or white. Urn A contains 50 red balls, 20 white balls, and 30 balls that are either red or white. Urns B and C each contain 30 red balls, 30 white balls, and 40 that are either red or white. We begin by drawing a ball from urn A. If the ball drawn from urn A is red, then we draw a second ball from urn B. If the ball drawn from urn A is white, then we draw a second ball from urn C. Once it is decided which urn to draw from, each ball in that urn is equally likely to be drawn. Let F be a variable whose values denote the color of the first ball drawn, and let S be a variable whose values denote the color of the second ball drawn. Let u be the number of red balls in urn B, and let v be the number of red balls in urn C.

The following proposition is true:

Proposition 20. *The variables F and S in Urn Example are completely independent if and only if $u = v$.*

This result fits with the idea that if an agent's imprecise joint credal set over two variables satisfies complete independence, then that agent ought to believe in the absence of a mechanistic connection between the phenomena represented by the two variables. In this case, if an agent's imprecise joint credal set over F and S satisfies complete independence, then that agent also ought to believe that $u = v$.

By contrast, we can give cases in which F and S are epistemically independent, but $u \neq v$. To illustrate, let us begin by defining marginal credal sets over F and S. Given that there are between 50 and 80 red balls in urn A, and between 30 and 70 red balls in urns B and C, these credal sets should be defined as follows.

- $\mathcal{K}[F = red] = [.5, .8]$
- $\mathcal{K}[F = white] = [.2, .5]$
- $\mathcal{K}[S = red] = [.3, .7]$
- $\mathcal{K}[S = white] = [.3, .7]$

[7] Note that Couso et al. (1999) is concerned with notions of independence in the imprecise context but not with imprecise Bayes nets.

In light of the definition of epistemic independence given above, we know that F and S are epistemically independent if and only if the following conditions hold:

- $\mathcal{K}[F = red|S = red] \subseteq [.5, .8]$
- $\mathcal{K}[F = red|S = white] \subseteq [.5, .8]$
- $\mathcal{K}[F = white|S = red] \subseteq [.2, .5]$
- $\mathcal{K}[F = white|S = white] \subseteq [.2, .5]$
- $\mathcal{K}[S = red|F = red] \subseteq [.3, .7]$
- $\mathcal{K}[S = red|F = white] \subseteq [.3, .7]$
- $\mathcal{K}[S = white|F = red] \subseteq [.3, .7]$
- $\mathcal{K}[S = white|F = white] \subseteq [.3, .7]$

That is, the conditional credal sets defined over each variable, given each possible value of the other variable, must be a subset of the marginal credal sets defined over that same variable.

There are conditional probability distributions that are consistent with the constraints listed above but are inconsistent with the assumption that $u = v$, assuming that all balls in a given urn are equally likely to be drawn from that urn. Suppose that there are 68 red balls in urn A, 69 red balls in urn B, and 37 red balls in urn C. This yields the following conditional probability assignments:

- $p(F = red|S = red) \approx .79$
- $p(F = red|S = white) \approx .51$
- $p(F = white|S = red) \approx .21$
- $p(F = white|S = white) \approx .49$
- $p(S = red|F = red) = .69$
- $p(S = red|F = white) = .37$
- $p(S = white|F = red) = .31$

- $p(S = white|F = white) = .63$

Clearly, conditional credal sets containing these conditional probabilities could satisfy the constraints required for epistemic independence. Just as clearly, these conditional probabilities are inconsistent with a lack of mechanistic connection between the first and second ball drawn; drawing a red ball first makes it much more likely that a red ball will be drawn second. Thus, taken pointwise, the credal sets described above allow for violations of complete independence. However, on a setwise appraisal, they satisfy epistemic independence.

This example shows that the sort of evidence that warrants the assignment of conditional credal sets over variables such that those variables are completely independent is different from the sort of evidence that warrants the assignment of conditional credal sets over variables such that those variables are epistemically independent. If the evidence available to an agent warrants the assignment of a joint credal set over two variables such that those two variables are completely independent, then that agent ought to believe that no objective, mechanistic connection exists between the two variables. In the Urn Example, this would mean that the agent ought to believe that the number of red balls in urn B is equal to the number of red balls in urn C, although the agent may not know how many red balls are in each urn. By contrast, if the evidence available to an agent warrants the assignment of conditional credal sets over two variables such that those two variables are completely but not epistemically independent, then this evidence is not decisive as to whether the number of red balls in urn B is equal to the number of red balls in urn C. All that an agent can conclude on the basis of this evidence is that learning the color of the first ball does not change the set of gambles that they would accept with respect to the color of the second ball. This distinction between evidence for complete independence and evidence for epistemic independence has important implications for the extent to which these independence concepts can be used to formulate adequacy conditions for the causal interpretation of an IBN.

6.5 Problems for an Imprecise Version of CMC and Faithfulness

I argue that neither complete nor epistemic independence can be neatly "plugged into" CMC and Faithfulness so as to allow for the causal interpretation of imprecise Bayes nets. To see why, let us begin by defining two possible versions of CMC, using each of the two independence concepts considered here:

Complete CMC: For any variable X in $\mathcal{I}$, X is completely independent of its non-descendants, conditional on its parents.

Epistemic CMC: For any variable X in $\mathcal{I}$, X is epistemically independent of its non-descendants, conditional on its parents.

Similarly, we can define the following versions of the Faithfulness condition:

Complete Faithfulness: For any variable X in $\mathcal{I}$, X is completely independent of its only non-descendants, conditional on its parents.

Epistemic Faithfulness: For any variable X in $\mathcal{I}$, X is epistemically independent of its only non-descendants, conditional on its parents.

Next, let us adopt the following necessary and sufficient condition for an IBN to have a causal interpretation:

Causal Interpretation Condition: An IBN $\mathcal{I} = \langle \mathcal{V}, \mathcal{E}, \mathcal{K} \rangle$ can be interpreted causally if and only if for every $p(\cdot) \in \mathcal{K}$, the Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ satisfies the precise versions of CMC and Faithfulness.

It turns out that graphs that satisfy any consistent combination of Complete CMC, Complete Faithfulness, Epistemic CMC, and Epistemic Faithfulness can still violate the Causal Interpretation Condition. Thus, neither set of adequacy conditions is sufficient for the causal interpretation of an IBN.

To see why, suppose that X and Y are variables in an IBN $\mathcal{I} = \langle \mathcal{V}, \mathcal{E}, \mathcal{K} \rangle$, and that they are epistemically but not completely independent conditional on their parents. If $\mathcal{I}$ satisfies Complete CMC and Complete Faithfulness, then there must be a directed path in either direction between X and Y. This directed path may simply be an edge between X and Y, such that the two variables are adjacent. However, $\mathcal{I}$ does not satisfy the Causal Interpretation Condition. Since X and Y are epistemically but not completely independent conditional on their parents, there may be some precise distribution $p_1(\cdot) \in \mathcal{K}$ such that if X is a descendent of Y, then Y is independent of X conditional on Y's parents, and if Y is a descendant of X, then X is independent of Y conditional on X's parents, where independence is understood in the precise sense. Under either of these circumstances, the graph $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p_1(\cdot) \rangle$ would violate the precise version of Faithfulness, since removing an edge on the directed path between X and Y would not create a subgraph that violates the precise version of CMC. Thus, $\mathcal{N}$ violates Minimality and herefore Faithfulness, implying that $\mathcal{I}$ violates the Causal Interpretation Condition. This shows that Complete CMC and

Complete Faithfulness are not sufficient conditions for an IBN to satisfy the Causal Interpretation Condition.

Next, consider an IBN $\mathcal{I}^- = \langle \mathcal{V}, \mathcal{E}^-, \mathcal{K} \rangle$, with variables X and Y that are epistemically but not completely independent, conditional on their parents. If $\mathcal{I}^-$ satisfies Epistemic CMC and Epistemic Faithfulness, then there must not be a directed path between X and Y, since the existence of such a path would violate Epistemic Faithfulness. Indeed, the only difference between $\mathcal{I}$ and $\mathcal{I}^-$ is that the former contains a directed path between X and Y, while the latter does not. It turns out that $\mathcal{I}^-$ also does not satisfy the Causal Interpretation Condition. Since X and Y are only epistemically independent, there is some $p_2(\cdot) \in \mathcal{K}$ such that X and Y are *not* independent conditional on their parents, in the precise sense, and yet there is no directed path between them. The graph $\mathcal{N}^- = \langle \mathcal{V}, \mathcal{E}^-, p_2(\cdot) \rangle$ would violate the precise version of CMC, so that $\mathcal{I}^-$ violates the Causal Interpretation Condition. Thus, Epistemic CMC and Epistemic Faithfulness are not a sufficient set of conditions for an IBN to satisfy the Causal Interpretation Condition. One can check that the same result holds if we suppose that $\mathcal{I}^-$ satisfies Epistemic CMC and Complete Faithfulness. Therefore, this mixed-strength combination of adequacy conditions also does not suffice to satisfy the Causal Interpretation Condition.

For the sake of completeness, it is worth noting that if there are variables X and Y in an IBN that are epistemically but not completely independent conditional on their parents, then that graph cannot satisfy Complete CMC and Epistemic Faithfulness. This is for the straightforward reason that when two variables in an IBN are epistemically but not completely independent conditional on their parents, Complete CMC is logically inconsistent with Epistemic Faithfulness. To see this, one need only note that since X and Y are epistemically but not completely independent, Complete CMC requires that there be a path between them but Epistemic Faithfulness requires that there not be such a path. Thus, the IBN cannot satisfy both conditions.

The upshot of this discussion is as follows. To interpret an IBN causally, we want it to be the case that every precise distribution in the credal set supports a causal interpretation of the graph. However, the possibility of epistemically but not completely independent variables is in conflict with this desideratum, no matter how we formulate imprecise versions of CMC and Faithfulness. To put this point slightly differently, if an IBN satisfies Complete CMC and contains two variables X and Y that are epistemically but not completely independent, then an edge between those variables may be a *false positive*. That is, the two variables may be adjacent, indicating a direct mechanistic connection between them, even though the joint credal set over the graph is consistent with some precise joint distribution according to which there is no direct mechanistic connection between X and Y. On the other hand, if an IBN satisfies Epistemic CMC and has two variables X and Y that are epistemically

but not completely independent, the *lack* of an edge between those variables may be a *false negative*. That is, the two variables are non-adjacent, indicating that there is no direct mechanistic connection between them, even though the joint credal set over the graph is consistent with some precise joint distribution according to which there *is* a direct mechanistic connection between X and Y.

The possibility of these kinds of false positives and false negatives injects considerable ambiguity into the causal interpretation of an IBN. Recall that in a precise Bayes net, if two variables are adjacent and there is no direct mechanistic connection between them, or if two variables are not adjacent and there is a mechanistic connection between them, then either there is a failure of causal sufficiency or the joint probability distribution is incorrect. By contrast, the results given above show that the following four claims can all consistently hold of an IBN: 1) two variables X and Y are adjacent, 2) there is no mechanistic connection between X and Y, 3) the variable set $\mathcal{V}$ is causally sufficient, and 4) the joint credal set $\mathcal{K}$ is accurate. Alternatively, the following four claims can also all be true of an IBN: 1*) two variables X and Y are not adjacent, 2*) there is a mechanistic connection between X and Y, 3) the variable set $\mathcal{V}$ is causally sufficient, and 4) the joint credal set $\mathcal{K}$ is accurate. The possible consistency of these claims shows an ambiguity in the causal interpretation of an IBN that is absent from the precise context.

6.6 Objections and Responses

6.6.1 Problems with the Causal Interpretation Condition

The obvious pressure point for my argument is the Causal Interpretation Condition. One could argue that it imposes too strong of a condition on the causal interpretation of an IBN, and that a weaker condition is needed. My response to this objection is to begin by showing that a much weaker condition on the causal interpretation of an IBN is clearly unworkable. Consider the following condition:

> **Weak Causal Interpretation Condition:** An IBN $\mathcal{I} = \langle \mathcal{V}, \mathcal{E}, \mathcal{K} \rangle$ can be interpreted causally if and only if there is *some* $p(\cdot) \in \mathcal{K}$ such that the Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ satisfies the precise versions of CMC and Faithfulness.

This condition is clearly too weak. Suppose that a credal set was defined over the variables in an IBN such that every combination of values in the graph was assigned to the set of probabilities between zero and one, exclusive. On the Weak Causal Interpretation Condition, such a graph would necessarily have a causal interpretation,

even though it seems to express an extreme level of ignorance about the dependence relationships between the variables in its target system.

However, one could put forward a condition that establishes a middle ground between my condition and the much weaker condition given above. That is, one could argue that an IBN $\mathcal{I} = \langle \mathcal{V}, \mathcal{E}, \mathcal{K} \rangle$ can be interpreted causally even if there is some $p(\cdot) \in \mathcal{K}$ such that the Bayes net $\mathcal{N} = \langle \mathcal{V}, \mathcal{E}, p(\cdot) \rangle$ does not satisfy the precise versions of CMC and Faithfulness. Any such argument would require further restrictions on the set of precise Bayes nets that are consistent with $\mathcal{I}$ and yet do not satisfy the precise versions of CMC and Faithfulness. While such restrictions may be well-justified given the context in which an imprecise Bayes net is deployed, there are no obvious *epistemic* reasons for adopting any such restriction. For instance, one might hold that an IBN can be interpreted causally as long as none of the precise Bayes nets with which it is consistent contain contain any edges that are possible false positives (possible false negatives are permitted). Such a position would be akin to adopting Epistemic CMC and Faithfulness as adequacy conditions for the causal interpretation of an IBN. Justifying this move would require an argument in favor of the claim that false negatives are somehow less pernicious than false positives when putting forward causal hypotheses. Surely such an argument would depend crucially on ethical or pragmatic premises, rather than purely epistemic considerations. While I take no issue with such extra-epistemic arguments in principle, I take it that the question of whether an IBN has a valid causal interpretation is itself a purely epistemic question that ought not to be answered by recourse to pragmatic or ethical considerations.

6.6.2 Eliminating Epistemic Independence

A different response to my argument could proceed as follows. The problems that I have raised with respect to a causal interpretation of an IBN only rear their heads once we consider that, in an imprecise Bayes Net, variables can be epistemically but not completely independent. This suggests an obvious, if decidedly *ad hoc*, way of positing adequacy conditions for the causal interpretation of an IBN. Let us posit that an IBN $\mathcal{I}$ can be interpreted causally if and only if it satisfies three conditions. The first two conditions are Complete CMC and Complete Faithfulness. The third condition is the following:

> **Complete Independence Only:** No two variables X and Y in $\mathcal{I}$ are epistemically independent but not completely independent.

It should be clear, given the discussion above, that these three conditions allow for the causal interpretation of an IBN while avoiding worries about false positives with respect to the existence of direct causal relations between variables.

While this restriction would permit the causal interpretation of an IBN, it is important to acknowledge that such a restriction significantly limits our ability to build a causal model from data. As I have argued above, the most accurate probabilistic representation of scientists' understanding of a system, given the available data, may be imprecise. It is conceivable that such a representation may imply that some random variables in the graph are epistemically but not completely independent. According to the constraints proposed in the previous paragraph, such a model would be disqualified from use in making causal inferences about the system under study. Thus, many reasonable representations of a system using imprecise probabilities would fail to justify any causal hypothesis.

This state of affairs stands in contrast to the situation facing the causal modeller who uses precise probabilities. Assuming causal sufficiency, one can draw causal conclusions from any precise joint probability distribution over the variables in $\mathcal{V}$. For instance, it may be that all of the variables in $\mathcal{V}$ are probabilistically independent (in the precise sense), in which case one would conclude that none of the variables are causally related. Alternatively, it may be that the joint probability distribution over $\mathcal{V}$ is consistent with a Bayes net from which various causal hypotheses may be derived. Indeed, the joint probability distribution over $\mathcal{V}$ may be consistent with several possible Bayes nets, each providing different representations of causal structure. The possibilities are highly variegated; my point is only that, on the assumption of causal sufficiency, any precise joint probability distribution over a variable set can be used in a model that shows either a causal relationship between two variables, or a lack thereof. By contrast, if we grant that Complete Independence Only is an adequacy condition for the causal interpretation of an IBN, then there are many joint credal sets that cannot be used to construct *any* causal model, because they are defined over variables that are epistemically but not completely independent. Thus, adopting Complete Independence Only would not ameliorate the problem for Imprecise Bayes Nets described above; rather, to adopt this condition would be to accept the problem as unsolvable.

This contrast between the power of imprecise and precise probabilistic models to represent causal structure is not especially surprising. The basic upshot is that when a model becomes less mathematically precise, it also loses some representational capacity with respect to the causal structure of that system. This is not to say that such a model is worse, on the whole, than a more precise representation of the same target system. It is just that whatever representational virtue the imprecise model has (e.g. that the model accurately represents the inherent ambiguity of

the available evidence about some system) comes at the expense of the model's capacity to represent the system's causal features. As an analogy, one could liken an IBN to an impressionist painting of a landscape and a Bayes net that uses precise probabilities to a photo-realistic painting of the same landscape. The former may have certain representational advantages over the latter—the impressionist style may convey more accurately the phenomenology of seeing the landscape—but these advantages have some costs with regard to the representational capacity of the painting—we cannot use the impressionist painting to make reliable inferences about the comparative heights of the various hills, as we might be able to when looking at the photo-realist painting.

6.6.3 Metaphysical Objections to Imprecise Causal Modelling

Finally, one might object from a different direction and claim that problems arise for a causal interpretation of IBNs before we even consider issues of independence between random variables in an imprecise model. Rather, the objection goes, imprecision is fundamentally incompatible with a causal interpretation of Bayes nets because the imprecision in an IBN represents an epistemic or subjective uncertainty about the values of variables and the relations between them, whereas probabilistic causal facts are grounded in objective chances. Implicit in this objection is the assumption that objective chances must be precise probabilities. Under this assumption, imprecise probabilistic attitudes towards the values of variables represent uncertainty as to the joint objective chance distribution over the graph. When faced with this kind of uncertainty, the objection continues, one should acknowledge that one lacks enough evidence to generate a causal hypothesis, even if there are no variables in one's model that are epistemically but not completely independent. Note that this claim is stronger than my central thesis, which is that, all else being equal, imprecision *threatens*, but does not necessarily eliminate, our ability to interpret a graphical model causally. If correct, this objection would render this entire chapter fundamentally misguided, since it would entail that we never should have been searching for adequacy conditions for the causal interpretation of an IBN in the first place.

There are two things to say in response to this objection. First, if it really is a requirement for the causal interpretation of a probabilistic graphical model that the joint probability distribution over the variables in the graph is an objective chance distribution, then few models actually used in the special sciences will have a causal interpretation. As List and Pivato (2015b) argue, most probabilities used in special science models reflect both the modeller's ignorance about some aspects of the target system and some observer-independent indeterminacy in the functioning of the

system. To claim that all such models are incapable of being used to formulate causal hypotheses, because the joint probability distribution represents some ignorance about outcomes, strikes me as a philosophical overreach into scientific practice and therefore out of keeping with a naturalistic approach to philosophy of science.

Further, the argument above hinges on the implicit and potentially unjustified assumption that the joint objective chance distribution over a set of variables must be a precise probability distribution. As Bradley (2016) argues, if objective chances are just taken to be those things in nature according to which agents ought to apportion their belief, then there is no requirement that beliefs formed in accordance with objective chances must be represented via precise probabilities. Indeed, it might be that an IBN is the *only* way of representing a target system in a way that accurately reflects the joint objective chance distribution according to the variables in the system. There may still be challenges with respect to the causal interpretation of the IBN, but these challenges would be due to the model's imprecision rather than its failure to represent objective indeterminacy in the target system.

6.7 Conclusion

This chapter has presented a formal model for Bayes nets that use imprecise rather than precise probabilities. I have shown that the adequacy conditions used in the causal interpretation of precise Bayes nets cannot be straightforwardly imported into an imprecise context. The upshot of this discussion is that imprecise Bayes nets, while they may have many other representational and practical virtues, often cannot be taken to represent causal structure. More generally, we can conclude that, all other factors being equal, precision is a virtue when it comes to the use of probabilistic models to represent the causal structure of the world. However, as should be clear from the arguments above, my claim is not that imprecision necessarily obviates the possibility of a causal interpretation of a graph. Rather, I conclude that while causal modelling with IBNs is possible, there are worrying ambiguities in the causal interpretation of an IBN that are absent from the precise context.

6.8 Appendix

6.8.1 Proof of Proposition 20

Proof. We begin by assuming that F and S are completely independent. This means that for any probability distribution in any joint credal set over F and S, the following holds (r and w stand for red and white):

$$p(F = r, S = r) = p(F = r)p(S = r) \tag{6.4}$$

Using the general definition of joint probability, the law of total probability, and some algebra, the equation can be re-written as follows:

$$\begin{aligned} p(S = r|F = r)p(F = r) = p(F = r)\big(&p(S = r|F = r)p(F = r) \\ &+ \; p(S = r|F = w)(1 - p(F = r))\big) \end{aligned} \tag{6.5}$$

$$\begin{aligned} p(S = r|F = r)p(F = r) = p(S = r|F = r)p(F = r)^2 + &p(S = r|F = w)p(F = r) \\ - \; p(S = r|F = w)p(F = r)^2 & \end{aligned} \tag{6.6}$$

Assuming that all probabilities are between zero and one exclusive, this equation holds if and only if $p(S = r|F = r) = p(S = r|F = w)$, i.e. the color of the first ball drawn does not make a difference to the probability that the second ball drawn is red. Under the conditions of the example, this is true if and only if urns B and C have the same number of red and white balls, i.e. $u = v$. Thus, F and S are completely independent if and only if $u = v$. □

7 Conclusion

There is much more work left to do. Over the course of this dissertation, I have defended a broadly pragmatic solution to the problem of how to choose the correct level of granularity for variables in probabilistic causal explanations. As noted in the introduction, this problem is a sub-problem of both the problem of granularity in scientific explanation, and the problem of variable choice in causal modelling. This leaves at least two immediate avenues for future work. Namely, one could investigate how the decision-theoretic arguments used in this dissertation could be applied to either the broader problem of granularity in scientific explanation, or the broader problem of variable choice in causal modelling.

One could also envision a fruitful research program in machine learning and applied philosophy of science that uses the coarsening methods proposed in Chapter 4 to make progress on various learning problems. Chalupka et al. (2015) apply their causal feature learning methods to problems involving computer vision, specifically the visual detection of causal patterns by computers. One could explore whether the approach to causal feature learning defended here could also be used as a framework for computer vision, and whether or not this framework would improve the performance of computer vision algorithms. Such a research program could then provide a basis for philosophical reflection on the pragmatic components of both perception in general and the detection of causal patterns in particular.

In addition to providing a basis for future work, this dissertation has also arrived at some definitive (though surely not uncontroversial) conclusions. To begin, I take myself to have given, in Chapter 1, a coherent argument for the claim that the Bayes nets approach to causal modelling supports an ecumenical response to various causal exclusion arguments. Next, I believe that my arguments in Chapter 2 show that there are some issues with the applicability of various Bayesian measures of explanatory power, at least when it comes to cases of scientific explanation across varying levels of granularity. Chapters 3 and 4 present a precise way of measuring the inherently pragmatic dimension of judgments of explanatory depth or goodness, which I take to be an ineliminable component of how scientists decide between better or worse explanations of a given phenomenon. Chapter 5 has offered an argument against what I take to be the strongest formal argument in favor of the claim that the probabilities used in the special sciences can be objective chances even

if the underlying microphysics is deterministic. Finally, Chapter 6 has made explicit some of the challenges that arise when attempting to represent causal structure using imprecise probability distributions.

However convincing my particular conclusions are, I hope that this dissertation has lived up to certain methodological ideals in the philosophy of science. Specifically, I have aimed throughout this dissertation to do philosophy of science in a way that makes concepts formally explicit and avoids conceptual vagueness, while also taking scientific practice seriously. In addition, I have aimed to draw, where appropriate, on discussions of causation and explanation that take place outside of the philosophical literature. I take this methodology to be consistent with a thoroughgoing naturalism. I also believe that my conclusions, with their emphasis on the practical aspects of scientific explanation and causal inference, can be traced to an extent to my focus on scientific practice. This approach can be contrasted with approaches to both explanation and causation that are more closely aligned with analytic metaphysics (see Woodward (2015, 2017) for an opinionated account of the contrast between these two approaches).

I am confident that there are errors in this dissertation (see Makinson (1965) for an elucidation of the logic that grounds this confidence). Although I had help from many people while writing this dissertation, I am solely responsible for all such errors.

Bibliography

Andersen, Holly. 2017. “Patterns, Information, and Causation”. *Journal of Philosophy* 114 (11): 592–622. (Cit. on pp. 68, 94).

Antonucci, Alessandro, Andrea Salvetti, and Marco Zaffalon. 2004. “Assessing debris flow hazard by credal nets”. In *Soft methodology and random Information Systems*, 125–132. Springer. (Cit. on pp. 153, 154).

Batterman, Robert W. 2001. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press. (Cit. on p. 77).

Bechlivanidis, Christos, David A Lagnado, Jeffrey C Zemla, and Steven Sloman. 2017. “Concreteness and abstraction in everyday explanation”. *Psychonomic bulletin & review* 24 (5): 1451–1464. (Cit. on p. 63).

Blackwell, David. 1951. “Comparison of Experiments”. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93–102. University of California Press. (Cit. on p. 87).

Bradley, Seamus. 2016. “Imprecise Probabilities”. In *The Stanford Encyclopedia of Philosophy*, Winter 2016, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University. (Cit. on pp. 15, 154, 172).

Campos, Cassio Polpo de, and Fabio Gagliardi Cozman. 2007. “Computing lower and upper expectations under epistemic independence”. *International Journal of Approximate Reasoning* 44 (3): 244–260. (Cit. on p. 162).

Cartwright, Nancy. 1979. “Causal Laws and Effective Strategies”. *Nous* 13 (4): 419–437. (Cit. on p. 145).

– . 1999. *The dappled world: A study of the boundaries of science*. Cambridge University Press. (Cit. on p. 105).

Chalupka, Krzysztof, Tobias Bischoff, Pietro Perona, and Frederick Eberhardt. 2016a. “Unsupervised Discovery of El Nino Using Causal Feature Learning on Microlevel Climate Data”. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 72–81. (Cit. on pp. 14, 99).

Chalupka, Krzysztof, Frederick Eberhardt, and Pietro Perona. 2017. “Causal feature learning: an overview”. *Behaviormetrika* 44 (1): 137–164. (Cit. on pp. 14, 99, 101, 117).

– . 2016b. “Multi-level cause-effect systems”. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 361–369. (Cit. on pp. 14, 99, 119).

Chalupka, Krzysztof, Pietro Perona, and Frederick Eberhardt. 2015. “Visual Causal Feature Learning”. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*. (Cit. on pp. 14, 99, 175).

Chuang, John S, Olivier Rivoire, and Stanislas Leibler. 2009. “Simpson’s paradox in a synthetic microbial system”. *Science* 323 (5911): 272–275. (Cit. on p. 145).

Clarke, Christopher. 2017. “How to define levels of explanation and evaluate their indispensability”. *Synthese* 194 (6): 2211–2231. (Cit. on pp. 6, 77, 93).

Cohen, Jonathan, and Craig Callender. 2009. “A Better Best System Account of Lawhood”. *Philosophical Studies* 145 (1): 1–34. (Cit. on p. 131).

Cohen, Michael P. 2016. “On Three Measures of Explanatory Power with Axiomatic Representations”. *British Journal for the Philosophy of Science* 67 (4): 1077–1089. (Cit. on p. 56).

Corani, G, A Antonucci, and M Zaffalon. 2012. “Bayesian networks with imprecise probabilities: Theory and application to classification”. *Data Mining: Foundations and Intelligent Paradigms* 23:49–93. (Cit. on pp. 15, 154, 158).

Couso, I, S Moral, and P Walley. 1999. “Examples of independence for imprecise probabilities.” In *Proceedings of the First Symposium on Imprecise Probabilities and Their Applications (ISIPTA), Ghent, Belgium*. (Cit. on pp. 155, 161–163).

Cozman, Fabio G. 2012. “Sets of probability distributions, independence, and convexity”. *Synthese* 186 (2): 577–600. (Cit. on pp. 155, 160).

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, Clarendon Press. (Cit. on p. 80).

Crupi, Vincenzo, and Katya Tentori. 2012. “A second look at the logic of explanatory power (with two novel representation theorems)”. *Philosophy of Science* 79 (3): 365–385. (Cit. on pp. 13, 53, 55).

Davidson, Donald. 1970. “Mental Events”. In *Experience and Theory*, ed. by L. Foster and J. W. Swanson, 79–101. Humanities Press. (Cit. on pp. 6, 17).

Dennett, Daniel C. 1991. “Real Patterns”. *Journal of Philosophy* 88 (1): 27–51. (Cit. on p. 94).

Eberhardt, Frederick. 2007. “Causation and intervention”. *Unpublished doctoral dissertation, Carnegie Mellon University*. (Cit. on p. 93).

Elkin, Lee, and Gregory Wheeler. 2018. "Resolving peer disagreements through imprecise probabilities". *Noûs* 52 (2): 260–278. (Cit. on p. 159).

Eronen, Markus. 2012. "Pluralistic Physicalism and the Causal Exclusion Argument". *European Journal for Philosophy of Science* 2 (2): 219–232. (Cit. on p. 18).

Eva, Benjamin, and Reuben Stern. Forthcoming. "Causal Explanatory Power". *The British Journal for the Philosophy of Science*. (Cit. on pp. 13, 29, 53, 56).

Fenton-Glynn, Luke. 2017. "A Proposed Probabilistic Extension of the Halpern and Pearl Definition of Cause'". *British Journal for the Philosophy of Science* 68 (4): 1061–1124. (Cit. on p. 25).

Fitelson, Branden. 2017. "Confirmation, Causation, and Simpson's Paradox". *Episteme* 14 (3): 297–309. (Cit. on p. 146).

Franklin-Hall, L. R. 2016. "High-Level Explanation and the Interventionist's 'Variables Problem'". *British Journal for the Philosophy of Science* 67 (2): 553–577. (Cit. on pp. 6, 42, 77, 85, 86, 93).

Garfinkel, Alan. 1981. *Forms of explanation: Rethinking the questions in social theory*. Yale University Press. (Cit. on pp. 62, 77).

Gebharter, Alexander. 2017. "Causal exclusion and causal Bayes nets". *Philosophy and Phenomenological Research* 95 (2): 353–375. (Cit. on pp. 13, 18, 37, 39).

Glennan, Stuart. 2002. "Rethinking mechanistic explanation". *Philosophy of Science* 69 (S3): S342–S353. (Cit. on p. 157).

Glynn, Luke. 2010. "Deterministic Chance". *British Journal for the Philosophy of Science* 61 (1): 51–80. (Cit. on p. 131).

Good, I. J. 1967. "On the Principle of Total Evidence". *British Journal for the Philosophy of Science* 17 (4): 319–321. (Cit. on p. 87).

Grice, H Paul. 1989. *Studies in the Way of Words*. Harvard University Press. (Cit. on p. 43).

Griffiths, P., A. Pocheville, B. Calcott, et al. 2015. "Measuring Causal Specificity". *Philosophy of Science* 82 (4): 529–55. (Cit. on p. 29).

Hájek, Alan. 2003. "What conditional probability could not be". *Synthese* 137 (3): 273–323. (Cit. on p. 105).

Hausman, DM, and J. Woodward. 1999. "Independence, Invariance and the Causal Markov Condition". *British Journal for the Philosophy of Science* 50 (4): 521–583. (Cit. on p. 21).

Hempel, Carl G, and Paul Oppenheim. 1948. "Studies in the Logic of Explanation". *Philosophy of Science* 15 (2): 135–175. (Cit. on pp. 5, 62).

Hitchcock, Christopher. 1999. "Contrastive Explanation and the Demons of Determinism". *British Journal for the Philosophy of Science* 50 (4): 585–612. (Cit. on p. 25).

– . 2012. “Events and times: a case study in means-ends metaphysics”. *Philosophical Studies* 160 (1): 79–96. (Cit. on p. 17).

– . 1996. “The Role of Contrast in Causal and Explanatory Claims”. *Synthese* 107 (3): 395–419. (Cit. on p. 25).

Hitchcock, Christopher, and James Woodward. 2003. “Explanatory Generalizations, Part II: Plumbing Explanatory Depth”. *Noûs* 37 (2): 181–199. (Cit. on pp. 6, 53, 60, 64, 82).

Hoefer, Carl. 2007. “The Third Way on Objective Probability: A Sceptic’s Guide to Objective Chance”. *Mind* 116 (463): 549–596. (Cit. on p. 131).

Hoover, Kevin D. 2001. *Causality in macroeconomics*. Cambridge University Press. (Cit. on p. 105).

Huang, Yimin, and Marco Valtorta. 2006. “Pearl’s calculus of intervention is complete”. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 217–224. (Cit. on pp. 27, 100, 104).

Jackson, Frank, and Philip Pettit. 1990a. “Causation and the Philosophy of Mind”. *Philosophy and Phenomenological Research* 50:195–214. (Cit. on p. 17).

– . 1988. “Functionalism and Broad Content”. *Mind* 97 (July): 318–400. (Cit. on pp. 6, 17).

– . 1992. “In Defense of Explanatory Ecumenism”. *Economics and Philosophy* 8 (1): 1–21. (Cit. on pp. 6, 17, 77).

– . 1990b. “Program Explanation: A General Perspective”. *Analysis* 50 (2): 107–17. (Cit. on pp. 6, 17).

Janzing, Dominik, and Bernhard Scholkopf. 2010. “Causal inference using the algorithmic Markov condition”. *IEEE Transactions on Information Theory* 56 (10): 5168–5194. (Cit. on p. 20).

Joyce, James M. 2010. “A defense of imprecise credences in inference and decision making”. *Philosophical perspectives* 24 (1): 281–323. (Cit. on p. 154).

– . 2009. “Accuracy and coherence: Prospects for an alethic epistemology of partial belief”. In *Degrees of Belief*, 263–297. Springer. (Cit. on p. 57).

Kim, Jaegwon. 1989. “Mechanism, Purpose, and Explanatory Exclusion”. *Philosophical Perspectives* 3:77–108. (Cit. on pp. 6, 18).

– . 2000. *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT press. (Cit. on pp. 6, 18).

Kinney, David. 2018. “Imprecise Bayesian networks as causal models”. *Information* 9 (9): 211. (Cit. on p. vii).

– . 2019. “On the Explanatory Depth and Pragmatic Value of Coarse-Grained, Probabilistic, Causal Explanations”. *Philosophy of Science* 86 (1): 145–167. (Cit. on p. vii).

Kitcher, Philip. 1981. “Explanatory Unification”. *Philosophy of Science* 48 (4): 507–531. (Cit. on pp. 77, 82).

Kolmogorov, Andrei N. 1933. *Foundations of the theory of probability: Second English Edition*. Ergebnisse Der Mathematik. (Cit. on p. 10).

– . 1963. “On tables of random numbers”. *Sankhyā: The Indian Journal of Statistics, Series A*: 369–376. (Cit. on p. 64).

Kolodny, Niko, and John MacFarlane. 2010. “Ifs and oughts”. *The Journal of philosophy* 107 (3): 115–143. (Cit. on pp. 131, 141).

Kotzen, Matt. 2013. “Conditional Oughts and Simpson’s Paradox.” *Unpublished manuscript*. (Cit. on p. 145).

Levi, Isaac. 1974. “On Indeterminate Probabilities”. *Journal of Philosophy* 71 (13): 391–418. (Cit. on pp. 158, 159).

– . 1980. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. MIT Press. (Cit. on pp. 158, 159, 161).

Lewis, David. 1986. “Causal explanation”. *Philosophical Papers* II:214–40. (Cit. on pp. 25, 42).

Lin, Hanti, and Jiji Zhang. 2018. “How to Tackle an Extremely Hard Learning Problem: Learning Causal Structures from Non-Experimental Data without the Faithfulness Assumption or the Like”. *arXiv preprint arXiv:1802.07051*. (Cit. on p. 105).

Lipton, Peter. 1990. “Contrastive Explanation”. *Royal Institute of Philosophy Supplement* 27:247–266. (Cit. on p. 25).

List, Christian. Forthcoming. “Levels: Descriptive, Explanatory, and Ontological”. *Nous*. (Cit. on p. 6).

List, Christian, and Peter Menzies. 2009. “Nonreductive Physicalism and the Limits of the Exclusion Principle”. *Journal of Philosophy* 106 (9): 475–502. (Cit. on pp. 13, 18, 40, 41, 83).

List, Christian, and Marcus Pivato. 2015a. “Dynamic and stochastic systems as a framework for metaphysics and the Philosophy of Science”. *arXiv preprint arXiv:1508.04195*. (Cit. on p. 132).

– . 2015b. “Emergent Chance”. *Philosophical Review* 124 (1): 119–152. (Cit. on pp. 15, 61, 62, 131, 135, 136, 138, 171).

Loewer, B. 2001. “Determinism and Chance”. *Studies in History and Philosophy of Science Part B* 32 (4): 609–620. (Cit. on p. 131).

Lyon, Aidan. 2011. “Deterministic Probability: Neither Chance nor Credence”. *Synthese* 182 (3): 413–432. (Cit. on p. 150).

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. “Thinking About Mechanisms”. *Philosophy of Science* 67 (1): 1–25. (Cit. on p. 157).

Makinson, David C. 1965. “The paradox of the preface”. *Analysis* 25 (6): 205–207. (Cit. on p. 176).

Malaterre, Christophe. 2011. “Making Sense of Downward Causation in Manipulationism (with Illustrations From Cancer Research)”. *History and Philosophy of the Life Sciences* 33 (4): 537–562. (Cit. on p. 80).

Malinas, Gary, and John Bigelow. 2016. “Simpson’s Paradox”. In *The Stanford Encyclopedia of Philosophy*, Fall 2016, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University. (Cit. on p. 145).

McDonald, Jennifer. ms. “The Proportionality of Common Sense Causal Claims”. (Cit. on p. 43).

Meek, Christopher. 1995. “Strong completeness and faithfulness in Bayesian networks”. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 411–418. Morgan Kaufmann Publishers Inc. (Cit. on pp. 105, 156).

Okamoto, Masashi. 1973. “Distinctness of the eigenvalues of a quadratic form in a multivariate sample”. *The Annals of Statistics*: 763–765. (Cit. on p. 104).

Pearl, Judea. 1994. “A probabilistic calculus of actions”. In *Uncertainty Proceedings 1994*, 454–462. Elsevier. (Cit. on p. 28).

– . 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. (Cit. on pp. 7, 17, 26, 28, 154).

Pearl, Judea, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons. (Cit. on p. 50).

Pearl, Judea, and Azaria Paz. 1985. *Graphoids: A graph-based logic for reasoning about relevance relations*. University of California (Los Angeles). Computer Science Department. (Cit. on p. 47).

Percival, Philip. 2000. “Lewis’s Dilemma of Explanation Under Indeterminism Exposed and Resolved”. *Mind* 109 (433): 39–66. (Cit. on p. 25).

Pettigrew, Richard. 2016. *Accuracy and the Laws of Credence*. Oxford University Press. (Cit. on p. 57).

Pettit, P. 2017. “The program model, difference-makers, and the exclusion problem”. In *Making a difference.* Ed. by H. Price H. Beebe C. Hitchcock. Oxford University Press. (Cit. on p. 17).

Pocheville, Arnaud, Paul E Griffiths, and Karola Stotz. 2017. “Comparing causes—an information-theoretic approach to specificity, proportionality and stability”. In *Proceedings of the 15th congress of logic, methodology and Philosophy of Science. College Publications, London*, 261–286. (Cit. on pp. 14, 29, 40, 78, 83, 101).

Polger, Thomas W., Lawrence A. Shapiro, and Reuben Stern. 2018. “In Defense of Interventionist Solutions to Exclusion”. *Studies in History and Philosophy of Science Part A* 68:51–57. (Cit. on p. 18).

Potochnik, Angela. 2015. “Causal patterns and adequate explanations”. *Philosophical Studies* 172 (5): 1163–1182. (Cit. on p. 94).

– . 2017. *Idealization and the Aims of Science*. University of Chicago Press. (Cit. on p. 94).

– . 2010. “Levels of explanation reconceived”. *Philosophy of Science* 77 (1): 59–72. (Cit. on p. 6).

Putnam, Hilary. 1979. *Philosophical Papers: Volume 2, Mind, Language and Reality*. Cambridge University Press. (Cit. on p. 77).

Reichenbach, Hans. 1956. *The Direction of Time*. Dover Publications. (Cit. on p. 21).

Resnik, Michael. 1987. *Choices: An Introduction to Decision Theory*. Univ of Minnesota Press. (Cit. on p. 87).

Salmon, Wesley C. 1998. *Causality and explanation*. Oxford University Press. (Cit. on p. 94).

Savage, Leonard J. 1954. *The Foundations of Statistics*. Wiley Publications in Statistics. (Cit. on p. 87).

Schaffer, J. 2007. “Deterministic Chance?” *British Journal for the Philosophy of Science* 58 (2): 113–140. (Cit. on p. 135).

Schaffer, Jonathan. 2016. “Grounding in the image of causation”. *Philosophical studies* 173 (1): 49–100. (Cit. on p. 39).

Schupbach, Jonah N., and Jan Sprenger. 2011. “The Logic of Explanatory Power”. *Philosophy of Science* 78 (1): 105–127. (Cit. on pp. 13, 53, 55, 58, 65).

Seidenfeld, Teddy, Mark J. Schervish, and Joseph B. Kadane. 2010. “Coherent choice functions under uncertainty”. *Synthese* 172 (1): 157–176. (Cit. on p. 159).

Shalizi, Cosma Rohilla, and James P Crutchfield. 2001. “Computational mechanics: Pattern and prediction, structure and simplicity”. *Journal of statistical physics* 104 (3-4): 817–879. (Cit. on p. 99).

Shalizi, Cosma Rohilla, and Cristopher Moore. 2003. “What is a macrostate? Subjective observations and objective dynamics”. *arXiv preprint cond-mat/0303625*. (Cit. on p. 150).

Shannon, Claude E., and Warren Weaver. 1949. *The Mathematical Theory of Communication*. University of Illinois Press. (Cit. on p. 83).

Shapiro, L., and E. Sober. 2012. “Against Proportionality”. *Analysis* 72 (1): 89–93. (Cit. on pp. 42, 93).

Shimizu, Shohei, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. “A linear non-Gaussian acyclic model for causal discovery”. *Journal of Machine Learning Research* 7 (Oct): 2003–2030. (Cit. on p. 20).

Simpson, E. H. 1951. “The Interpretation of Interaction in Contingency Tables”. *Journal of the Royal Statistical Society. Series B (Methodological)* 13 (2): 238–241. (Cit. on p. 145).

Sklar, Lawrence. 1993. *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge University Press. (Cit. on p. 77).

Sober, Elliott. 2010. “Evolutionary Theory and the Reality of Macro Probabilities”. In *The Place of Probability in Science*, ed. by Ellery Eells and James H. Fetzer, 133–60. Springer. (Cit. on p. 131).

Sober, Elliott, and David Sloan Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press. (Cit. on p. 145).

Spirtes, Peter, Clark Glymour, Scheines N., and Richard. 2000. *Causation, Prediction, and Search*. MIT Press: Cambridge. (Cit. on pp. 7, 17, 19–21, 26, 28, 105, 154, 157).

Spirtes, Peter, and Richard Scheines. 2004. “Causal Inference of Ambiguous Manipulations”. *Philosophy of Science* 71 (5): 833–845. (Cit. on p. 35).

Steel, Daniel. 2006. “Homogeneity, selection, and the faithfulness condition”. *Minds and Machines* 16 (3): 303–317. (Cit. on p. 105).

Stegmann, Ulrich. 2014. “Causal Control and Genetic Causation”. *Noûs* 48 (3): 450–465. (Cit. on p. 80).

Strevens, Michael. 2008. *Depth: An account of scientific explanation*. Harvard University Press. (Cit. on pp. 6, 53, 60, 77).

– . 2011. “Probability Out Of Determinism”. In *Probabilities in Physics*, ed. by Claus Beisbart and Stephan Hartmann, 339–364. Oxford University Press. (Cit. on p. 150).

Ströing, Pascal. 2018. “Data, Evidence, and Explanatory Power”. *Philosophy of Science* 85 (3): 422–441. (Cit. on pp. 13, 54, 68, 69).

van Fraassen, 1980. *The Scientific Image*. Oxford University Press. (Cit. on p. 94).

Verma, Thomas, and Judea Pearl. 1990a. “Causal networks: Semantics and expressiveness”. In *Machine Intelligence and Pattern Recognition*, 9:69–76. Elsevier. (Cit. on p. 156).

Verma, TS, and Judea Pearl. 1990b. “Equivalence and synthesis of causal models”. In *Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence*, 220–227. (Cit. on p. 47).

Walley, Peter. 1991. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall. (Cit. on pp. 159, 161).

Weatherson, Brian. 2012. “Explanation, Idealisation and the Goldilocks Problem”. *Philosophy and Phenomenological Research* 84 (2): 461–473. (Cit. on pp. 6, 77, 86, 93).

Weslake, Brad. 2010. “Explanatory Depth”. *Philosophy of Science* 77 (2): 273–294. (Cit. on pp. 6, 14, 34, 53, 60, 62, 77, 78, 81, 87, 91, 93).

– . 2013. “Proportionality, Contrast and Explanation”. *Australasian Journal of Philosophy* 91 (4): 785–797. (Cit. on pp. 77, 78, 80, 81).

Williamson, Jon. 2005. *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press. (Cit. on p. 22).

Wilson, Alastair. 2018. “Metaphysical causation”. *Noûs* 52 (4): 723–751. (Cit. on p. 39).

Woodward, James. 2010. “Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation”. *Biology and Philosophy* 25 (3): 287–318. (Cit. on pp. 77, 80, 91).

– . 2018. “Explanatory Autonomy: The Role of Proportionality, Stability, and Conditional Irrelevance”. *Synthese*: 1–29. (Cit. on p. 82).

– . 2015a. “Interventionism and Causal Exclusion”. *Philosophy and Phenomenological Research* 91 (2): 303–347. (Cit. on p. 18).

– . 2017. “Interventionism and the Missing Metaphysics”. *Metaphysics and the Philosophy of Science: New essays*: 193. (Cit. on p. 176).

– . 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. (Cit. on pp. 7, 8, 18, 24).

– . 2015b. “Methodology, ontology, and interventionism”. *Synthese* 192 (11): 3577–3599. (Cit. on p. 176).

– . 2016. “The Problem of Variable Choice”. *Synthese* 193 (4): 1047–1072. (Cit. on pp. 12, 77).

– . 2002. “What is a mechanism? A counterfactual account”. *Philosophy of Science* 69 (S3): S366–S377. (Cit. on p. 157).

Yablo, Stephen. 1992. “Mental Causation”. *Philosophical Review* 101 (2): 245–280. (Cit. on pp. 40, 80).

Yarkoni, Tal, and Jacob Westfall. 2017. “Choosing prediction over explanation in psychology: Lessons from machine learning”. *Perspectives on Psychological Science* 12 (6): 1100–1122. (Cit. on p. 65).

Ylikoski, Petri, and Jaakko Kuorikoski. 2010. “Dissecting Explanatory Power”. *Philosophical Studies* 148 (2): 201–219. (Cit. on pp. 82, 87, 91).

Yule, G Udny. 1903. “Notes on the theory of association of attributes in statistics”. *Biometrika* 2 (2): 121–134. (Cit. on p. 145).

Zhang, Jiji. 2012. “A comparison of three occam’s razors for markovian causal models”. *The British Journal for the Philosophy of Science* 64 (2): 423–448. (Cit. on p. 23).