ON CAUSAL INFERENCES IN THE HUMANITIES AND SOCIAL SCIENCES: ACTUAL CAUSATION

A Dissertation by

Jonathan Mark Livengood

B.A., B.S., Truman State University, 2004

M.A., University of Pittsburgh, 2011

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of
DOCTOR OF PHILOSOPHY

University of Pittsburgh

2011

ON CAUSAL INFERENCES IN THE HUMANITIES AND SOCIAL SCIENCES: ACTUAL CAUSATION

A Dissertation

by

Jonathan Mark Livengood

Defended

August 17, 2011

And approved as to style and content by

John Norton, Co-Chair _____

Peter Spirtes, Co-Chair _____

Robert Krafty, Member _____

Edouard Machery, Member _____

Sandy Mitchell, Member _____

ON CAUSAL INFERENCES IN THE HUMANITIES AND SOCIAL SCIENCES: ACTUAL CAUSATION

Jonathan Mark Livengood, PhD

University of Pittsburgh, 2011

The last forty years have seen an explosion of research directed at causation and causal inference. Statisticians developed techniques for drawing inferences about the likely effects of proposed interventions: techniques that have been applied most noticeably in social and life sciences. Computer scientists, economists, and methodologists merged graph theory and structural equation modeling in order to develop a mathematical formalism that underwrites automated search for causal structure from data. Analytic metaphysicians and philosophers of science produced an array of theories about the nature of causation and its relationship to scientific theory and practice.

Causal reasoning problems come in three varieties: effects-of-causes problems, causes-of-effects problems, and structure-learning or search problems. Causes-of-effects problems are the least well-understood of the three, in part because of confusion about exactly what problem is supposed to be solved. I claim that the problem everyone is implicitly trying to solve is the problem of identifying the actual cause(s) of a given effect, which I will call simply *the problem of actual causation*. My dissertation is a contribution to the search for a satisfying solution to the problem of actual causation.

Towards a satisfying solution to the problem of actual causation, I clarify the nature of the problem. I argue that the only serious treatment of the problem of actual causation in the

statistical literature fails because it confuses actual causation with simple difference-making. Current treatments of the problem of actual causation by philosophers and computer scientists are better but also ultimately unsatisfying. After pointing out that the best current theories fail to capture intuitions about some simple voting cases, I step back and ask a methodological question: how is the correct theory of actual causation to be discovered? I argue that intuition-fitting, whether by experimentation or by armchair, is misguided, and I recommend an alternative, pragmatic approach. I show by experiments that ordinary causal judgments are closely connected to broadly moral judgments, and I argue that actual causal inferences presuppose normative, not merely descriptive, information. I suggest that the way forward in solving the problem of actual causation is to focus on norms of proper functioning.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to take this opportunity to thank some people and institutions that have made this dissertation possible. I would especially like to thank Peter Spirtes for our discussions of causal reasoning problems and for many helpful reviews of my writing. I would also like to thank Edouard Machery for my introduction to experimental philosophy and Robert Krafty for his guidance on all things statistical. Thanks also to John Norton and Sandy Mitchell for their support and encouragement.

I would like to thank the American Council of Learned Societies for their financial support during the last year of writing this dissertation; it was much appreciated. Thanks to my wife, Kerrith Livengood, for her love and support over the course of writing this dissertation. Thanks also to my parents and the rest of my family for their optimism.

Special thanks to my good friend and sometimes collaborator, Justin Sytsma, for innumerable discussions, suggestions, and encouragements over the last three years. Thanks to Christopher Hitchcock, Jim Woodward, and Joshua Knobe for helpful comments on early drafts of some chapters in this dissertation. Thanks to my friend Balázs Gyenis for giving me a hard time during WIP talks. Thanks also to my colleagues Karen Zwier, Jonah Schupbach, Peter Distelzweig, Benny Goldberg, and Bryan Roberts for comments on earlier drafts and insightful conversations on demand. And generally, thanks to the other students, faculty, and staff associated with the department of History and Philosophy of Science at the University of Pittsburgh. If I have left out anyone to whom I owe a debt of gratitude, please accept my sincere apology.

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Increasingly, policy makers and pundits have expressed concerns about the weak performance of U.S. students on international tests of science and mathematics ability (Anderson 2010; Etter 2010; Will 2011). In 2009, among the 34 participating OECD nations, the U.S. tied with eleven other nations for 18th place in mathematics literacy and tied with twelve other nations for 13th place in science literacy (OECD 2010). What accounts for the poor performance of U.S. students on international science and mathematics tests, and what should be done to improve performance going forward?

Primed by the documentary *Waiting for Superman*, most of the recent discussion about education and education reform in the United States over the last year has centered on charter schools and teachers' unions (Peterson 2010; Mahler 2011; Ravitch 2011). Most of the education reformers answer the causal questions as follows: Student performance is inadequate because too many teachers are ineffective, and in order to improve performance, the educational system should make it easier to fire ineffective teachers and do a better job of rewarding effective teachers. The reformers draw on attempts by economists and education researchers to estimate *education production functions* (Aaronson, et al. 2007; Boyd, et al. 2008; Kane and Staiger 2008; and Buddin and Zamarro 2009). As I write this, more than a dozen U.S. states have passed education reform legislation eliminating collective bargaining, introducing merit pay, expanding charter schools, changing tenure rules, or some combination thereof (Banchero

2011; Barron 2011; Calefati 2011; Gabriel and Dillon 2011; Ghianni 2011; Mazzei 2011; Resmovits 2011a, 2011b).

Some critics of the new reformers argue that weak student performance is due to widespread poverty in the U.S., not poor teacher quality (Krashen 2010; McCabe 2010; Riddile 2010). McCabe and Riddile report an interesting observation about the PISA test scores. The United States has much more significant economic diversity from one school to the next, and that diversity is related to the PISA scores. McCabe and Riddile focus on reading scores. They point out that whereas the average reading score in the U.S. is 500, the average reading score for schools with fewer than ten percent of their students eligible for the federal free lunch program is 551, which is second only to Shanghai's 556. Taking free lunch eligibility as a proxy for poverty rate and comparing U.S. schools to other participating nations based on similarity of poverty level, the U.S. comes out at the top of every category. Poverty rate is statistically associated with PISA scores. McCabe and Riddile explicitly draw the causal conclusion: poverty causes poor test scores.

Krashen, both in news opinion pieces and also in peer-reviewed research articles, defends the claim that poverty causes low test scores, and he is not alone.[1] Krashen also suggests a mechanism: poor children have reduced access to books and other educational resources (Krashen 2004, 2010; Krashen, et al. 2010). Studies by other researchers have confirmed the inverse statistical relationship between poverty and test scores, and they have also extended that observation to mathematics scores (Payne and Biddle 1999; Neuman and Celano 2001; Berliner 2006, 2009; Burnett and Farkas 2009). Increased poverty is associated with lower standardized test scores. However, some researchers are skeptical that the relationship is causal (Turner 2000; Payne and Biddle 2000; Myers, et al. 2004). Disagreement about the causal relations leads the

[1] See Sirin 2005 for a review of the literature.

various parties in the education reform debate to different answers to the two questions: (1) what accounts for the poor performance of U.S. students on international tests and (2) what should be done to improve scores in the future? One side thinks the answers come back to the quality of students' teachers; the other side thinks the answers come back to the quality of students' lives.

Over the last forty years, statisticians have developed techniques for answering causal questions like, "Would such and so intervention improve test performance in the future?" which are called questions about the *effects of causes*. At the same time, economists, philosophers, and computer scientists have developed formally related techniques for answering causal questions like, "What is the causal structure involving poverty, academic achievement, and assorted other factors?" However, questions like, "What caused the poor performance of U.S. students on tests in 2009," which are called *causes of effects* questions by statisticians and *actual causation* questions by philosophers and computer scientists working with structural equation models, are still poorly understood. The difficulties with cause-of-effect problems go deep. Not only are there technical challenges involved in answering cause-of-effect questions, the very concept of an "actual cause" is unclear.

The only serious treatment of cause-of-effect questions in the statistical literature—Dawid (2000)—equates actual causation with simple difference-making. But the metaphysical literature on causation provides compelling reasons to think that actual causation is not the same thing as simple difference-making. Philosophers have developed numerous competing accounts of actual causation, even when one restricts attention to the structural equation modeling tradition. Moreover, philosophers do not agree about how to decide which of the competing accounts of actual causation is correct. Finally, no one has a clear idea what purpose cause-of-

effect reasoning serves, although everyone agrees it is fundamental to legal reasoning in criminal and tort cases!

My dissertation is a contribution to the search for an adequate account of actual causation. An adequate account of actual causation must have three parts: a metaphysical part, a logical part, and a pragmatic part. The metaphysical part gets at the nature of actual causation. What makes something an actual cause?[2] The logical part gets at the norms for drawing inferences about actual causation. Given an account of the nature of actual causation, we want to know what we can learn about actual causation and under what conditions. The pragmatic part gets at the purpose(s) of having an account of actual causation in the first place.

In addition to the account itself, we also need an adequate meta-theory. The dominant approach to theorizing about causation in the philosophical literature is a guess-and-check search constrained by the armchair intuitions of causal theorists. Some theorists have used experiments on folk intuitions to constrain the search, but using such experiments does not (by itself) change the method in any significant way. And in the case of causation, at least, augmenting the armchair Socratic meta-theory with experimentation does not produce an adequate meta-theory. I argue that a pragmatic, Euclidean meta-theory is better. Instead of trying to fit intuitions, the theorist first identifies the purpose(s) served by a concept of actual causation and then asks what theory best accomplishes the purpose(s) of the concept.

Here is how I will proceed. In Chapter 1, I discuss the dominant approach to causal inference in the statistical literature and an alternative approach recommended by Dawid (2000).

[2] Two other problems are often associated with the metaphysical problem. The first problem is the psychological problem of saying what the ordinary concept or concepts of actual causation are like insofar as such concepts exist. The second problem is the linguistic problem of specifying the semantics for ordinary discourse about actual causation. The psychological and linguistic problems are related insofar as one's concept(s) of actual causation together with pragmatic constraints give rise to ordinary discourse about actual causation. The psychological and semantic problems are related to the metaphysical problem insofar as one's concepts track the way actual causation really works, or to put it in a more pragmatic way, the psychological and semantic problems are related to the metaphysical problem insofar as they are fit for the work that we want from a concept of actual causation.

I formally distinguish cause-of-effect reasoning and effect-of-cause reasoning, and I discuss the relationship between causal reasoning and counterfactuals within the statistical framework. I argue that Dawid's approach to cause-of-effect reasoning is inadequate, since he incorrectly conflates actual causation and simple difference-making. I accept a desideratum for causal inference that Dawid defends: causal inferences should depend on as few arbitrary, untestable assumptions as possible. I suggest that graphical, structural equation approaches to actual causation are better able to satisfy that desideratum.

In Chapter 2, I describe three basic varieties of causal reasoning problem from the perspective of the graphical, structural equation modeling approach to causation. I distinguish structural causation, which underwrites effects-of-causes reasoning, from actual causation, which underwrites causes-of-effects reasoning. I clarify the role of random variables in structural causal models and the relationship between random variables in a causal model and properties. I further note that the actual-structural distinction has sometimes been confused with the singular-generic distinction, and I show how they are different.

In Chapter 3, I consider several current accounts of actual causation coming out of the structural equation approach. I argue that simple voting scenarios are prima facie counter-examples to all of these accounts. In Chapter 4, I consider an empirical demand that (many) philosophers accept with respect to causation: the folk attribution desideratum (FAD). According to the FAD, a theory of (actual) causation is in bad trouble if it does not correctly predict the causal attributions that ordinary people make. Given that background, I present some simple experiments undermining a promising account of actual causation offered by Hitchcock (2007). Finally, in Chapter 5, I step back and discuss meta-theory for normative accounts of actual causation. I describe two forms of Socratic meta-theory: Cartesian rationalism and

deflationary rationalism. I then describe the pragmatic, Euclidean meta-theory that I favor. I discuss the relationship between the various meta-theories and experimental philosophy. I then discuss some specific difficulties that experimentalists face when researching causation.

CHAPTER ONE

COUNTERFACTUALS AND CAUSATION IN STATISTICS

The dominant approach to causal inference in statistics today is the Neyman-Rubin framework, named after Jerzy Neyman (1990 [1923])—who is usually credited with originating the framework—and Donald Rubin (1974, 1976, 1977, 1978, 1997, 2006)—who developed the details of the framework starting in the 1970s. The Neyman-Rubin framework is counterfactual in character, making use of what are called *potential outcomes*. However, counterfactuals have commitments that cannot be empirically tested. Untestable assumptions about counterfactuals (at least, assumptions untestable up front—more on this later) turn out to have practical implications for causal inference. Moreover, in the Neyman-Rubin framework, inferences that depend on untestable assumptions are not clearly distinguished from those that do not. Hence, users of the counterfactual approach to causation are apt to be misled. Hence, Dawid (2000) argues that the Neyman-Rubin framework with its commitment to counterfactuals ought to be given up (for the most part) in favor of a decision theoretic approach.

In the present chapter, I explore some issues surrounding Dawid's rejection and replacement of the counterfactual approach to causal inference. Following Holland (1986), Dawid distinguishes two broad types of causal inference: inference about the effects of causes (EoC) and inference about the causes of effects (CoE). Dawid claims to show how to dispense with counterfactual assumptions for (some) typical inferences about the effects of causes. (Extending his examples to more complicated cases is not straightforward.) However, he accepts

that counterfactual assumptions are unavoidable for making inferences about the causes of effects.

In this chapter, I review some of the discussion surrounding Dawid's paper in order to motivate a structural equations approach to causes-of-effects inferences. In Section 1.1, I discuss Dawid's philosophy of statistics, which provide desiderata for an adequate theory of cause-of-effects inferences. In Section 1.2, I describe three distinct varieties of causal reasoning problem: effects-of-causes problems, causes-of-effects problems, and structure-learning problems. Since Dawid only discusses the first two varieties, the rest of the chapter concentrates on those. In Section 1.3, I set out the philosophical and statistical machinery for effects-of-causes inferences both in the Neyman-Rubin framework, which appeals to counterfactuals, and also in Dawid's decision-theoretic framework, which does not. In Section 1.4, I turn to causes-of-effects inferences. I argue that Dawid and his discussants ought to be talking about actual causation but are in fact talking about something else: simple difference-making. In Section 1.5, I reply to some generic criticisms of causes-of-effects inferences raised by Shafer (2000), one of Dawid's discussants.

1.1 DAWID'S PHILOSOPHY

Dawid's complaint against the Neyman-Rubin framework is that it makes causal inferences depend on assumptions that are untestable in principle. The basis of his concern is broadly verificationist or instrumentalist.[3] Dawid distinguishes theories or quantities that have testable consequences (at least in principle) and theories or quantities that do not. The former are such that one can say what difference the truth of the theory or the value of the quantity would make

[3] Dawid claims that his approach is "grounded in a Popperian philosophy, in which the meaningfulness of a purportedly scientific theory … is related to the implications it has for what is or could be observed" (408).

to some data that one could, in principle, acquire. Dawid calls these kinds of theories or quantities *scientific*. Those theories or quantities that cannot be connected to data, even in principle, he calls *metaphysical*.[4] Dawid writes:

> I argue that counterfactual theories are essentially metaphysical. This in itself might not be automatic grounds for rejection of such a theory, if the causal inferences that it led to were unaffected by the metaphysical assumptions embodied in it. Unfortunately, this is not so, and the answers that the approach delivers to its inferential questions are seen, on closer analysis, to be dependent on the validity of assumptions that are entirely untestable, even in principle. This can lead to distorted understandings and undesirable practical consequences. … The general message of this article is that inferences based on counterfactual assumptions and models are generally unhelpful and frequently plain misleading. Alternative approaches can avoid these problems, while continuing to address meaningful causal questions. (408-409)

Dawid claims that some causal inferences in the Neyman-Rubin framework require untestable assumptions. These inferences have practical consequences, but nothing in the formal apparatus of the Neyman-Rubin framework alerts the user to the fact that he or she is making an inference that depends on untestable, arbitrary assumptions. Hence, the counterfactual approach is apt to mislead its users.

To what extent relying on arbitrary assumptions is problematic is left somewhat unclear. Dawid claims (see the previous quotation) that some inferences in the counterfactual approach rely on assumptions that are untestable in principle but nevertheless have practical consequences. However, those practical consequences are not clearly articulated. In a response article, Pearl (2000) proposes the following dilemma for Dawid's warning against using counterfactuals:

> [Dawid's warning] is either empty or self-contradictory. If one's conclusions have no practical consequences, then their sensitivity to invalid assumptions is totally harmless, and Dawid's warning is empty. If, on the other hand, one's conclusions do have practical consequences, then their sensitivity to assumptions automatically makes those assumptions testable, and Dawid's warning turns contradictory. (429)

Dawid appears to miss the point of Pearl's dilemma, replying:

[4] Dawid's choice of terminology is not at all provocative!

> I am in total agreement with this … as long as attention is confined to such testable aspects, no problem arises. My point is that the models I criticize also have *untestable* implications, and that (unless one takes great care) it is all too easy to use them to make "inferences" that are sensitive to purely arbitrary choices that may be made for ingredients in these models. (445)

I will return to this dispute in Section 1.2.3 below. For now, the take-home message is this: Dawid wants to restrict causal inferences to those that require non-arbitrary, testable assumptions and that make the assumptions easy to spot in the formalism. I agree with these desiderata, but I do not think that Dawid's decision-theoretic approach best satisfies those desiderata.

1.2 THREE VARIETIES OF CAUSAL REASONING PROBLEM

Dawid points out that "several different problems of causal inference … are often conflated" (408). One distinction that Dawid thinks is especially important is that between inferences about the effects of causes on the one hand and inferences about the causes of effects on the other—a distinction he attributes to Holland (1986). Inferences about the effects of causes begin with a cause or something assumed to be a cause for the purpose of the analysis. The typical effects-of-causes problem is to estimate the result (effect) of a specified intervention or treatment (cause). For example, one might be interested to know how much science test scores would be improved (or degraded) by an intervention that reduced relative poverty in the United States from 20% to 15%. By contrast, inferences about the causes of effects begin with a known outcome.[5] For example, one might want to know why something has the value is does, e.g. why the mean 2009 PISA science score for U.S. students was 502, or one might want to know why something changed in the way it did, e.g. why the mean PISA science score for U.S. students changed from

[5] One might ask about purely hypothetical cases or about cases that have not yet occurred but which might occur in the future. However, when one considers such cases, one treats the outcome (effect) as known for the purposes of the analysis. And the linguistic construction is also more complicated. For example, one might ask, "In such and so circumstances, if this were to have occurred (or had occurred), would that be (or have been) a cause of the occurrence?"

489 in 2006 to 502 in 2009. Given a known outcome, one wants to identify its cause(s).[6] Dawid illustrates the distinction by contrasting two questions about the relationship between headaches and aspirin:

(EC) I have a headache. Will it help if I take an aspirin?
(CE) My headache has gone. Was it because I took aspirin?

The question posed in (EC) is about the effect of a specified cause. The question posed in (CE) is about the cause of an observed effect. An answer to the question in (EC) would describe a kind of law and a rule for action. An answer to the question in (CE) would provide a kind of explanation.

In his *System of Logic*, Mill distinguishes between cases in which one asks what circumstances brought about a given outcome (causes of effects) and cases in which one asks what follows from given actions or occurrences (effects of causes). Mill argues that the plurality of causes of complex events, which are typical, raises serious problems for observational inference that a specific action or occurrence caused a given outcome. In Book III, Chapter x, Section 8 of his *Logic*, Mill writes:

> The inapplicability of the method of simple observation to ascertain the conditions of effects dependent on many concurring causes, being thus recognized; we shall next inquire whether any greater benefit can be expected … by directly trying different combinations of causes, either artificially produced or found in nature, and taking notice what is their effect: as, for example, by actually trying the effect of mercury [on health], in as many different circumstances as possible. This method differs from the one which we have just examined, in turning our attention directly to the causes or agents, instead of turning it to the effect, recovery from the disease. And since, as a general rule, the effects of causes are far more accessible to our study than the causes of effects, it is natural to think that this method has a much better chance of proving successful than the former. (449)

[6] An interesting variant that has not been addressed in the statistical literature and has mostly been intentionally ignored by philosophers is the problem of identifying the most important or influential cause or causes of an observed outcome.

Later, Mill uses the plurality of causes to challenge Bacon's theory of scientific method. In Book V, Chapter iii, Section 7 of his *Logic*, Mill writes:

> When [Bacon] is inquiring into what he terms the *forma calidi aut frigidi gravis aut levis*, *sicci aut himidi* and the like, he never for an instant doubts that there is some one thing, some invariable condition or set of conditions, which is present in all cases of heat, or cold, or whatever other phenomenon he is considering; the only difficulty being to find what it is; which accordingly he tries to do by a process of elimination, rejecting or excluding, by negative instances, whatever is not the *forma* or cause, in order to arrive at what is. But, that this *forma* or cause is *one* thing, and that it is the same in all hot objects, he has no more doubt of, than another person has that there is always some cause *or other*. In the present state of knowledge it could not be necessary, even if we had not already treated so fully of the question to point out how widely this supposition is at variance with the truth. ... Bacon was seeking for what did not exist. The phenomenon of which he sought for the one cause has oftenest no cause at all, and when it has, depends (as far as hitherto ascertained) on an unassignable variety of distinct causes.
>
> And on this rock every one must split, who represents to himself as the first and fundamental problem of science to ascertain what is the cause of a given effect, rather than what are the effects of a given cause. It was shown, in an early stage of our inquiry into the nature of Induction, how much more ample are the resources which science commands for the latter than for the former inquiry, since it is upon the latter only that we can throw any direct light by means of experiment; the power of artificially producing an effect, implying a previous knowledge of at least one of its causes. If we discover the causes of effects, it is generally by having previously discovered the effects of causes: the greatest skill in devising crucial instances for the former purpose may only end, as Bacon's physical inquiries did, in no result at all. (763-764)

Contemporary statisticians agree with Mill in thinking that effects-of-causes inferences are more tractable than causes-of-effects inferences. Holland (1986) introduces the distinction between effects of causes and causes of effects into the contemporary literature.[7] Holland argues that statistics applies first and foremost to inferences about the effects of causes. He writes:

> The emphasis [in this paper] will be on *measuring the effects of causes* because this seems to be the place where statistics, which is concerned with measurement, has contributions to make. It is my opinion that an emphasis on the effects of causes is, in itself, an important consequence of bringing statistical reasoning to bear on the analysis of causation and directly opposes more traditional analyses of causation. (945)

[7] Although Holland (1986) has a section on Mill, he does not attribute the distinction between effects of causes and causes of effects to Mill.

In a later paper, Holland (1993) argues that inferences about the causes of effects are parasitic on inferences about the effects of causes (implicitly agreeing with Mill). Dawid (2007) goes further: "CoE-type questions are generally more of intellectual than practical interest—with the notable exception of legal proceedings to determine liability. They are also much more problematic, both philosophically and methodologically" (17). Gelman (2010) offers a more recent perspective, though he largely agrees with the other writers. Though acute philosophers and statisticians have made enormous strides in causal reasoning over the last thirty years, approaches to causes-of-effects inferences are still under-developed.

In addition to effects-of-causes problems and causes-of-effects problems are structure-learning problems. In both effects-of-causes and causes-of-effects problems one is concerned with the causal relationship between specific values of pairs of random variables. By contrast, in structure-learning problems, one is concerned with the causal relationships between (or among) the random variables in some collection $\boldsymbol{V}$ of random variables. Such relations are structural. The collection of (structural) causal relationships between (or among) the random variables in a collection $\boldsymbol{V}$ of random variables is the causal structure over $\boldsymbol{V}$. Causal structure will be described in much greater detail in Chapter 2. For now, "causal structure" should be understood as roughly synonymous with "causal laws."

1.3 EFFECTS OF CAUSES

In the present section, I set out the philosophical and statistical machinery for drawing effects-of-causes inferences. I first describe the counterfactual framework in Sections 1.3.1 and 1.3.2. Then in Section 1.3.3, I consider the issue of testability. Finally, I contrast the counterfactual framework with Dawid's decision-theoretic framework in Section 1.3.4.

1.3.1 Philosophical Counterfactuals

One way of thinking about the history of philosophical logic is as an attempt to formalize the ordinary language *conditional*, which in English is written, "If …, then ---." Beginning in the second half of the nineteenth century, philosophers became interested in a special kind of conditional, the *counterfactual* conditional. Counterfactual conditionals (or simply counterfactuals) have constructions, like, "If *y* had been …, then *x* would have been ---," or "If *y* were …, then *x* would be ---."[8]

In the 1960s and 1970s, philosophers developed a semantic characterization of counterfactuals in terms of comparative similarity between possible worlds. The most influential account of counterfactuals is that of Lewis (1973a). Let w_a, w_b, and w_c be possible worlds. The world w_a is said to be *closer* to the world w_b than it is to the world w_c if w_a is overall more similar to w_b than it is to w_c.[9] In Lewis' semantics for counterfactuals, we say that if *c* were the case, then *e* would be the case (denoted $c \mathrel{\square\!\!\rightarrow} e$) iff either (i) there exists no possible world in which *c* occurs, or (ii) there exists a possible world in which both *c* and *e* occur that is closer to the actual world than any world in which *c* and $\sim e$ occur.

Lewis makes use of his theory of counterfactuals to define a relation of *causal dependence* as follows.[10] An event *e* causally depends on another event *c* iff the following two

[8] Philosophers usually lump all subjunctive conditionals into the category of counterfactuals. Strictly speaking, a counterfactual has an antecedent that is false; however, the antecedent of a subjunctive like, "If *x* were to be …, then *y* would be ---," might not be contrary to fact. Rather, the subjunctive sentence might be hypothetical. Counterfactuals, like, "If I had gone to sleep earlier, then I wouldn't be so tired," are subjunctive conditionals in which the antecedent is false. In this case, I didn't go to sleep earlier. Hypotheticals like, "If I were to buy my wife a wide-format printer, then she would be very happy," are subjunctive conditionals in which the truth-value of the antecedent is indeterminate. In this case, it isn't determinate whether I will buy my wife a printer or not. The terminology of "hypotheticals" is due to Dawid (2006), who notes that hypothetical conditionals are testable in a way that (strictly) counterfactual conditionals are not.

[9] Lewis takes comparative similarity between worlds to be a very vague primitive notion. However, he does discuss the overall similarity relation in several places (Lewis 1973a, 91-95; Lewis 1973b; and Lewis 1979).

[10] Although Rubin was apparently unaware of Lewis when he was explicitly developing the potential outcomes framework in the 1970s, by the mid-1980s, the statisticians begin to take Lewis' possible world semantics on board (see Holland 1986 and discussion).

counterfactuals hold: (i) $c \Box\!\!\rightarrow e$, and (ii) $\sim c \Box\!\!\rightarrow \sim e$. To this theory, Lewis had to make an immediate qualification. The counterfactuals in Lewis' account of causal dependence must be *non-backtracking*. A counterfactual is said to be *backtracking* if its truth requires that the past be different in order to accommodate a difference in the present. Here is Lewis' example:

> Jim and Jack quarreled yesterday, and Jack is still hopping mad. We conclude that if Jim asked Jack for help today, Jack would not help him. But wait: Jim is a prideful fellow. He never would ask for help after such a quarrel; if Jim were to ask Jack for help today, there would have to have been no quarrel yesterday. In that case Jack would be his usual generous self. So if Jim asked Jack for help today, Jack would help him after all. (1979, 456)

The claim that if Jim were to ask Jack for help today, there would have to have been no quarrel yesterday is a backtracking counterfactual. Lewis claims that ordinary counterfactuals are non-backtracking. He argues that backtracking counterfactuals have a different syntactic structure than non-backtracking counterfactuals. For example, backtracking counterfactuals have constructions like, "If *y* had been ___, then *x* would have to have been ___," or, "If *y* had been ___, then *x* would have had to have been ___."

For Lewis, when we want to decide whether an event *e* causally depends on another event *c*, we should imagine a miracle whereby the event *c* is replaced by ~*c* and the rest of the world is made to accommodate ~*c* by changes to events *later than* c. Philosophers occasionally debate whether causal relations or counterfactual conditional relations are more primitive: whether one *should* be analyzed in terms of the other or vice versa. Lewis thought that causation should be analyzed in terms of counterfactuals. Those who think that counterfactuals ought to be analyzed in terms of causation think that positing the two ideas of a miracle and of a time-order on events sneaks causal concepts into Lewis' analysis and makes it circular.

1.3.2 Statistical Counterfactuals

Statisticians take a formally very different approach to counterfactuals, but the main idea is strikingly similar. Let U be a set called the universe of discourse or the population. Each element u of U is called a unit. For simplicity, I will treat U as discrete and index the units of U with the subscript $i \in \mathbb{N}$. Let T and Y be random variables over the units. Call T the *treatment variable* and Y the *response variable*. Following Dawid, I restrict attention to single-intervention, single-response systems, where T is binary with possible values t and c for "treatment" and "control," respectively. The Neyman-Rubin framework assumes that for all i, $T(u_i)$ could have been either t or c. In other words, the property measured by T is manipulable such that each unit could have been assigned either to the treatment group or to the control group. The counterfactual framework postulates two new variables: the *potential outcome* variables, $Y_{T=t}$ and $Y_{T=c}$. The expression $Y_{T=t}(u_i)$ represents the value that the response variable Y would have taken for unit u_i if that unit had been assigned to the treatment group. Similarly, $Y_{T=c}(u_i)$ represents the value that the response variable Y would have taken for unit u_i if that unit had been assigned to the control group.

Before being assigned a value for the treatment variable, the potential outcome variables represent the value that the response variable would take under different hypothetical assignments. *After* being assigned to a treatment or control group, one potential outcome variable effectively becomes the *actual response variable* and the other becomes a *counterfactual response variable*. The counterfactual response variable $Y_{T=j}$ gives the answer to the question, "If the value of the treatment variable had been j, what value would the response variable Y have taken?" Like Lewis' counterfactuals, the counterfactuals in the Neyman-Rubin framework are non-backtracking.

Letting $t = 1$ and $c = 0$, the response variable is related to the treatment and potential outcome variables by the *consistency relation*,

(CR) $$Y = TY_{T=t} + (1 - T)Y_{T=c}.$$

In virtue of the way the potential outcome variables are defined, the values for the pair $Y_{T=t}$ and $Y_{T=c}$ cannot both be observed for the same unit. Holland (1986) calls the fact that $Y_{T=t}$ and $Y_{T=c}$ cannot both be observed for the same unit the *fundamental problem of causal inference*. Define the individual treatment effect (ITE) with respect to unit u_i as

(ITE) $$Y_{T=t}(u_i) - Y_{T=c}(u_i)$$

If the ITE is non-zero, then the treatment variable T is said to cause the response variable Y, and the effect of treatment $T = t$ (as opposed to treatment $T = c$) with respect to unit u_i is equal to the ITE.[11] The fundamental problem of causal inference may now be restated as follows: the individual treatment effect is unobservable in principle. No one can measure the effect of a given treatment (in contrast to some other treatment, usually the control) for an individual unit.

If the units are uniform with respect to possible treatments, then the ITE may be estimated by comparing different units, since if the units are uniform, then

(U) $$Y_{T=t}(u_i) - Y_{T=c}(u_i) = Y_{T=t}(u_i) - Y_{T=c}(u_j), \text{ for } i \neq j.$$

As long as $T(u_i) = t$ and $T(u_j) = c$ are the actual treatment values, the ITE is observable when the units are uniform with respect to treatment. Alternatively, if the units have a stable response to treatment over time and the effect of treatment is transient, then the ITE may be estimated by measuring the resting value of the response variable for the unit (which is equivalent to the value given assignment to the control group) and then applying treatment and measuring the treated

[11] In the event that the treatment variable has more than two possible values, the ITE is defined contrastively treatment by treatment.

response. In many research areas, especially in the hard sciences, assumptions of uniformity, stability, and transience appear plausible. In other research areas, they do not.

By contrast, the statistical solution to the fundamental problem of causal inference is to give up on the ITE itself and focus on its expectation instead. Estimating the expected value of the ITE—called the *average treatment effect* (ATE) or sometimes the *average causal effect* (ACE)—is the prototypical effects-of-causes problem.[12]

(ATE) $\mathrm{E}(Y_{T=t} - Y_{T=c})$

By the linearity of expectation, $\mathrm{E}(Y_{T=t} - Y_{T=c}) = \mathrm{E}(Y_{T=t}) - \mathrm{E}(Y_{T=c})$. If the treatment variable T is independent of the pair of potential outcome variables $\langle Y_{T=t}, Y_{T=c} \rangle$, then

(RA) $\mathrm{E}(Y_{T=t}) - \mathrm{E}(Y_{T=c}) = \mathrm{E}(Y_{T=t} \mid T = t) - \mathrm{E}(Y_{T=c} \mid T = c)$,

which is usually considered to hold in experimental setups where treatment is randomly assigned, since random assignment of treatment is assumed to eliminate confounding (at least in the large-sample limit). Under the *randomization assumption*, equation RA holds whenever treatment is randomly assigned to some collection of units (usually by applying some physical randomization device).[13] From the consistency relation, it follows that

(ICR) $\mathrm{E}(Y_{T=t} \mid T = t) - \mathrm{E}(Y_{T=c} \mid T = c) = \mathrm{E}(Y \mid T = t) - \mathrm{E}(Y \mid T = c)$.

The right-hand side of equation ICR is observable and can be estimated by comparing the response measured for the treatment group and the response measured for the control group. Hence, when the treatment is randomly assigned, the ATE is straightforwardly estimable. However, as we will see in a moment, the quality of the estimate is not estimable without untestable assumptions.

[12] Dawid uses the latter terminology. I find "average treatment effect" to be a bit clearer and will use it instead.
[13] Of course, equation RA might hold without randomization of treatment assignments.

If T is not independent of the potential outcome variables $Y_{T=t}$ and $Y_{T=c}$, then the ATE cannot be correctly estimated by the difference between the observed values for the units in the treatment group and the observed values for the units in the control group. For example, let T measure whether an individual went to a private high school ($T = t$) or a public high school ($T = c$), and let Y measure whether an individual scored above 1400 on the SAT ($Y = 1$) or not ($Y = 0$). Suppose that wealthier families are more likely to pay for effective SAT tutoring for their children.[14] In that case, the estimate of the ATE obtained by comparing the observed values for the units in the treatment group and the observed values for the units in the control group will be systematically larger than the true ATE. (Alternative stories could make the estimated ATE come out as systematically smaller than the true ATE.)

Dawid presents the following model for consideration. Suppose that the two potential outcome variables $Y_{T=t}(u_i)$ and $Y_{T=t}(u_i)$ follow a bivariate normal model consistent with equation

(1.1) $$Y_{T=j}(u_i) = \theta_j + \beta(u_i) + \gamma_j(u_i),$$

where j may take on the value t or c. The term θ_j is the mean of $Y_{T=j}$ and does not depend on the unit being considered. The terms $\beta(u_i)$ and $\gamma_j(u_i)$ are independent normal random variables with mean zero and variances $\phi_\beta = \rho\phi_Y$ and $\phi_\gamma = (1-\rho)\phi_Y$, respectively. Variable β represents a random unit effect, since it has a possibly unique value for each u_i and is insensitive to changes in the value of the treatment variable. Variable γ_j represents a unit-treatment interaction, since its value depends on both the unit and the assigned treatment. The variances of Y, β, and γ_j are related according to equations

(1.2) $$\phi_Y = \phi_\beta + \phi_\gamma$$

and

[14] By saying that the tutoring is effective, I mean to assert that tutoring *causes* greater proficiency on the SAT.

(1.3) $$\rho = \frac{\phi_\beta}{\phi_\beta + \phi_\gamma}.$$

Assuming the model in (1.1), the ITE is given by

(1.4) $$\tau(u_i) = \theta_t + \gamma_t(u_i) - [\theta_c + \gamma_c(u_i)] = \theta_t - \theta_c + \gamma_t(u_i) - \gamma_c(u_i).$$

The β terms for $Y_{T=t}(u_i)$ and $Y_{T=c}(u_i)$ cancel. The ATE for the model in (1.1) is $\theta_t - \theta_c$, since the β and γ terms have distributions with means equal to zero. Denote the ATE by $\tau = \theta_t - \theta_c$. Then, the distribution of the ITE for the model in (1.1) is normal with mean τ and variance $2\phi_\gamma$. Let $\phi_\tau = 2\phi_\gamma$. Then, $\tau(u_i) \sim N(\tau, \phi_\tau)$. While the mean of the distribution is estimable under the assumption that RA holds, the variance of the distribution is not. One may estimate τ with $\hat{\tau} = \hat{\theta}_t - \hat{\theta}_c$, where $\hat{\theta}_t$ and $\hat{\theta}_c$ are the sample means of the response variable Y for the treatment and control groups, respectively. However, the variance ϕ_τ cannot be estimated without making some assumption about the joint distribution of $Y_{T=t}$ and $Y_{T=c}$. Specifically, in order to estimate ϕ_τ, one needs to make some assumption about the value of ρ. Therefore, in order to evaluate the quality of the estimate of τ, one needs to make some assumption about the value of ρ. However, the value of ρ for the joint distribution of $Y_{T=t}$ and $Y_{T=c}$ is unobservable in principle (without some assumptions, like uniformity).

The typical choice—going back to Neyman (1923)—is to take $\rho = 1$. This assumption is called *treatment-unit additivity* (TUA). Treatment-unit additivity is equivalent to the assumption that $\tau(u_i)$ is the same for all i. Given TUA, the estimate of τ is perfect. Alternatively, one might take $\rho = 0$, in which case $\phi_\tau = 2\phi_\gamma = 2\phi_Y$. Strictly speaking, there are an uncountable number of additional alternative choices for the value of ρ, and none of them is testable.

1.3.3 Testability

According to Dawid, assumptions about the value of ρ sometimes have practical consequences in the counterfactual framework, even though the value of ρ cannot be estimated from data. Pearl objects that if two supposedly untestable assumptions have different practical consequences, then by comparing the utility of those consequences, one may test the assumptions themselves relative to one another—the assumption that has the most utility survives the test. The disagreement raises two questions. First, do assumptions like TUA have practical consequences as Dawid claims? Second, if assumptions like TUA have practical consequences, are they testable?

Take the questions in order: Do assumptions like TUA have practical consequences as Dawid claims? The answer is yes. Different assumptions about the value of ρ result in different credible intervals for the ATE, and hence, different assumptions about the value of ρ sometimes matter for whether or not one should assert that there is a genuine effect of treatment at all in a given case. Two statisticians who make different choices for the value of ρ might disagree about whether two variables are causally related, even if they are given the same data. Both the way one understands the world and the way one acts on the basis of the data depend on assumptions about the value of ρ.

Assumptions about the value of ρ (sometimes) have practical consequences at least insofar as they bear on the actions one takes. But is Pearl right when he says that so long as different assumptions about the value of ρ lead to different actions, we test to see which assumption about the value of ρ is correct? The answer is yes and no. We cannot test the value of ρ on the same data. However, we can test to see which assumption is correct by collecting a larger sample that reduces the uncertainty in our estimate for τ.

In the case that different assumptions about ρ lead to different decisions, enough follow-up sampling should resolve the disagreement. After disagreement has been resolved, the "correct" value of ρ will still be unobservable; however, as Pearl suggests, once the disagreement about whether or not there is a real causal connection is resolved, the value of ρ does not make any practical difference. Now, one might try to resolve the ambiguity before initially sampling. By taking a guess at the value of τ, one may calculate the worst-case confidence interval that would be required in order to make a decision not dependent on the value of ρ. Working backwards, one may then calculate the sample size needed, just as is done with power calculations.

1.3.4 Decision Theory

Dawid's decision theoretic approach to inferences about the effects of causes is very simple. Suppose that the potential outcome variables $Y_{T=t}$ and $Y_{T=c}$ have marginal distributions P_t and P_c, respectively. Dawid assumes that the marginal distributions have been consistently estimated from the data. The problem is now a decision problem about a new unit, u_0, which was not part of the data used to estimate the marginal distributions P_t and P_c. Specifically, we are to decide whether or not to apply treatment to u_0. And we measure the worth of our decision with a loss function $L(\cdot)$ of the (so-far unobserved) response Y. Dawid draws the decision tree reproduced in Figure 1.1.

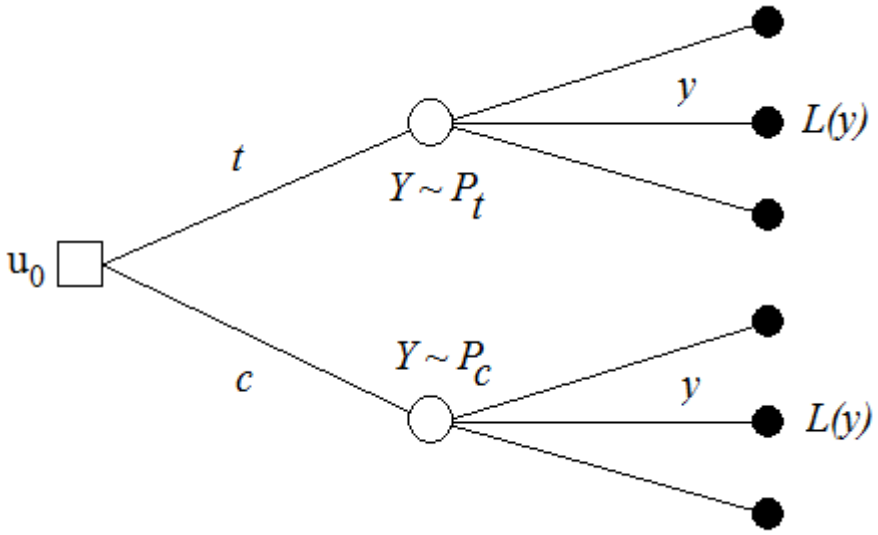

Figure 1.1: Decision Tree for Unit u_0

One should act (according to Bayesian decision theory) so as to minimize expected loss. Hence, the decision with respect to u_0 depends only on the marginal distributions P_t and P_c.

Dawid considers two statisticians, S1 and S2, who agree about the observable data but make different assumptions about the value of ρ. In the counterfactual framework, the two statisticians might draw different causal inferences; however, in the decision theoretic framework their disagreement makes no practical difference. He writes:

> It simply does not matter that S2 believes that the time for a headache to disappear if aspirin is taken will be *exactly* 10 minutes less than if it is not taken, whereas S1 regards the difference of these times as uncertain, although again with expectation 10 minutes; there is no way in which such differences in beliefs can affect the decision problem. (412)

Effectively, then, Dawid's decision theoretic approach encourages researchers to ignore the variance of the distribution of the ITE when making causal decisions. The upshot, oddly enough, is that for all practical purposes Dawid agrees with those statisticians who work in the counterfactual framework and endorse TUA.

Aside from the practical agreement between Dawid and those who endorse TUA in the counterfactual framework, Dawid's decision theoretic approach is *deceptively* simple. Most of the work is done by the assumption that the marginal distributions P_t and P_c of $Y_{T=t}$ and $Y_{T=c}$, respectively, may be consistently estimated from the data. In the decision theoretic framework, one assumes that the *observed* response values in the treatment and control cases are the same as the *potential* response values. Formally, one assumes both that the equation

(RA-t) $\quad \mathrm{E}(Y_{T=t}) = \mathrm{E}(Y_{T=t} \mid T = t)$

and that the equation

(RA-c) $\quad \mathrm{E}(Y_{T=c}) = \mathrm{E}(Y_{T=c} \mid T = c)$

hold. But if there is any confounding, then this assumption will not be correct. The assumption that the marginal distributions are consistently estimable from the data entails the assumption

(RA) $\quad \mathrm{E}(Y_{T=t}) - \mathrm{E}(Y_{T=c}) = \mathrm{E}(Y_{T=t} \mid T = t) - \mathrm{E}(Y_{T=c} \mid T = c),$

which appeared earlier in Section 1.2.2. In the counterfactual framework, the distinction between $\mathrm{E}(Y_{T=j} \mid T = j)$ and $\mathrm{E}(Y_{T=j})$ is relatively easy to state and to understand, and so, it is relatively easy to understand what it means to say that the two quantities are equal or to say that RA holds, which is slightly weaker. Moreover, the formalism makes the inferential import of the difference between $\mathrm{E}(Y_{T=j} \mid T = j)$ and $\mathrm{E}(Y_{T=j})$ clear, since one cannot estimate the ATE without making assuming that RA holds.

However, Dawid's approach hides the difference between $\mathrm{E}(Y_{T=j} \mid T = j)$ and $\mathrm{E}(Y_{T=j})$, and so, it hides the fact that the two are being assumed to be equal. Nothing in the formalism alerts the user that a serious assumption is being made when the observed (conditional) distributions are used in a decision. Moreover, the assumptions RA-t and RA-c (which entail RA) are not testable. Since the marginal distributions of the potential response variables are unobservable in

principle, we cannot compare those distributions to the observed conditional distributions. Insofar as Dawid is concerned that the formalism make its assumptions clear so that users do not unwittingly draw inferences with practical consequences from untestable assumptions, the fact that the decision theoretic approach hides the difference between $E(Y_{T=j} \mid T = j)$ and $E(Y_{T=j})$ should be regarded as a defect. In other words, we cannot test whether all confounding has been controlled or eliminated.[15]

1.4 CAUSES OF EFFECTS

As noted above, many statisticians and philosophers maintain that inferences about the causes of effects are both scientifically uninteresting and inaccessible to statistical methods. Recall Dawid's claim that questions about the causes of effects "are generally more of intellectual than practical interest—with the notable exception of legal proceedings to determine liability" (2007, 17). Shafer (2000, 441) goes so far as to say that questions about the causes of effects are "silly." And Gelman (personal communication) remarks, "I take the pretty much orthodox position in statistics that it's pretty meaningless to estimate the cause of an effect. When I think about statistical inference for causality, it's always estimating the effect of a cause."

The attitude towards cause-of-effect reasoning expressed by these statisticians is both surprising and disappointing, since searching for the causes of effects is, I contend, an important concern of applied science. For example, diagnosticians working in hospitals, trouble-shooters working in information technology, epidemiologists tracing the causes of an outbreak, and engineers doing fault analysis in a wide range of different areas are all implicitly interested in the

[15] In his own description of the counterfactual approach, Dawid does not remark on RA, and he claims that randomized assignment of treatment is inessential to the discussion (see 409). Hence, I cannot tell whether he is aware of the issue raised here or not.

causes of effects. The interest does not guarantee that answers are to be had, but I suspect that more answers would be forthcoming if statisticians were less dismissive.

1.4.1 The Statisticians' Problem

Dawid characterizes the inferential problem as follows. Suppose you are given all of the information as in the decision-theoretic version of the effects-of-causes problem. In addition, you know the value of the response variable for unit u_0. The problem is to decide whether or not $T(u_0) = j$ caused $Y(u_0) = y$, where y is the actual value of the response variable Y for unit u_0. Dawid actually assumes that the question has the form: given that the unit was assigned to the treatment group, what would have happened had the unit been assigned to the control group instead? Hence, he claims that in order to make this decision, it is required to compare the observed value y of $Y(u_0)$ to the counterfactual quantity $Y_{T=t}(u_0) = y_t$. He writes, "Equivalently, inference about the individual causal effect [ITE] $\tau(u_0) = y - Y_{T=t}(u_0)$ is required" (417). Though Dawid does not say this explicitly, the basic idea appears to be that a treatment $T(u_0) = j$ may be said to have caused an outcome $Y(u_0) = y$ if and only if the ITE $\tau(u_0) = y - Y_{T=t}(u_0)$ is different from zero. Hence, Dawid appears to be committed to the view that the causes of an effect are all and only the difference-makers in the actual case.

In order to keep his exposition simple, Dawid again considers the bivariate normal model consistent with equation

$$Y_{T=j}(u_i) = \theta_j + \beta(u_i) + \gamma_j(u_i), \tag{1.1}$$

where j may take on the value t or c. The conditional distribution of $\tau(u_0) = Y_{T=t}(u_0) - Y_{T=c}(u_0)$ given $Y_{T=t}(u_0) = y$ is normal with mean

$$\lambda = E[\tau(u_0) \mid Y_{T=t}(u_0) = y] = y - \theta_c - \rho(y - \theta_t) \tag{1.5}$$

and variance

(1.6) $\phi_\lambda = (1-\rho^2)\phi_Y$.

As in the effects-of-causes problem, θ_t, θ_c, and ϕ_Y can be estimated from the data. Unlike in the effects-of-causes problem, the mean of the target distribution cannot be estimated from the data alone. Hence, in the case of inference about the causes of effects, all of the properties of the target distribution depend on untestable assumptions about the joint distribution of the potential outcome variables, specifically the value of ρ.

Under TUA ($\rho = 1$), the mean of the target distribution is just the ATE, $\lambda = \theta_t - \theta_c$, and the variance is zero. So, given TUA, inference about the causes of effects works exactly like inference about the effects of causes. Assuming that $\rho = 0$, the mean is $\lambda = y - \theta_c$ and the variance is $\phi_\lambda = \phi_Y$. Assuming that $\rho = -1$, the mean is $\lambda = 2y - \theta_t - \theta_c$ and the variance is zero (again). As Dawid points out, inference about λ will be insensitive to untestable assumptions about ρ only when $Y_{T=t}(u_0) = y$ is sufficiently close to the mean of $Y_{T=t}$. Similarly, inference about ϕ_λ will be insensitive to untestable assumptions about ρ only when ϕ_Y is small. The good news is that those are things we can know on the basis of the data. Hence, we may be able to know in some cases that inferences about the causes of effects are legitimate, despite not being able to identify the value of ρ. Moreover, definite bounds may be given for the values of λ and ϕ_λ, again raising the possibility that the uncertainty about ρ will not matter in some cases where y is not close to the mean of $Y_{T=t}$ and ϕ_Y is not small.

1.4.2 Dawid's Solution: Concomitants

Dawid despairs of completely eliminating ambiguity in inferences about the causes of effects. "However," he writes, "some progress toward reducing this [ambiguity] may be possible if one can probe more deeply into the hidden workings of the units, by observing suitable additional variables" (418). Dawid shows how, by measuring concomitant variables assumed to be

unaffected by the application of treatment to the units, one may place significant constraints on the freedom of choice one has about non-identifiable parameters.

Again, assume that model (1.1) is correct. Suppose that experimenters have found that conditional on $K(u_i) = k$, $Y_{T=j}(u_i)$ is normally distributed with mean $\theta_j + k$ and (residual) variance ψ_K. Let ϕ_K denote the variance of K and also let $\phi_K = \text{cov}(K, Y_{T=c}) = \text{cov}(K, Y_{T=t})$. Define $\psi_0 = \phi_Y = \phi_K + \psi_K$. Then the partial correlation of $Y_{T=c}$ and $Y_{T=t}$ given K is given by

$$\rho_{ct.K} = 1 - (1 - \rho)\frac{\psi_0}{\psi_K}, \tag{1.7}$$

where, as before, ρ cannot be estimated from the data. Dawid considers two possibilities: (1) that K is observed with respect to the unit u_0, and (2) that K is not observed with respect to the unit u_0, though it is observed with respect to the other units. When $K(u_0) = k$ is observed, equations (1.5) and (1.6) may be replaced with

$$\lambda_K = y - \theta_c - k - \rho_{ct.K}(y - \theta_t - k) \tag{1.8}$$

and

$$\phi_{\lambda K} = (1 - \rho_{ct.K}^2)\psi_K, \tag{1.9}$$

respectively. The standard deviation of the $(y - \theta_t - k)$ term in (1.8) is strictly smaller than the corresponding term in (1.5). Similarly, with the information about $K(u_0)$, the variance is bounded above by $\psi_K < \phi_Y$. Hence, the measurement of a covariate with respect to u_0 makes the mean less sensitive to arbitrary choices about ρ (now $\rho_{ct.K}$, instead) and the estimate more exact, since the variance is closer to zero. Even when $K(u_0)$ is not observed, we can still use (1.7) to restrict the possible values of ρ to $[1 - 2\frac{\psi_K}{\psi_0}, 1]$, instead of their original range [-1, 1].

1.4.3 Wasserman and Pearl's Formulation of the Statistician's Problem

The problem as Dawid formulates it appears to have no probabilistic content in the sense that what is being estimated is whether the choice of treatment made an actual difference to the value of the response variable. The question for Dawid is whether the ITE is different from zero for unit u_0. However, in their response articles, Wasserman (2000) and Pearl (2000) rewrite Dawid's example of a causes-of-effects query in terms of conditional probability (or conditional expectation) as

(CE*) $\quad P(Y_{T=c} = 1 \mid T = t, Y = 0).$

In words, CE* is the conditional probability that the response Y would have been 1 (e.g., I would now have a headache) if T had been set to the control value (e.g., if I had been made not to take an aspirin) given that the actual treatment was $T = t$ (e.g., given that I actually took an aspirin) and the actual response to the assigned treatment was $Y = 0$ (e.g., given that I do not now have a headache).

On Wasserman and Pearl's framing, the interest is not so much in the discovery or identification of the causes of a given effect but rather in a retrospective effects-of-causes inference.[16] That is, one is trying to see how much of a difference the suspected cause was expected to make to the known effect. That framing is unfortunate, since it encourages a blurring together of (1) the claim that $T(u_0) = j$ caused $Y(u_0) = y$, and (2) the claim that $T(u_0) = j$ was a difference-maker with respect to $Y(u_0) = y$. Dawid conflates (1) and (2). Wasserman and Pearl further confuse (1) and (2) with the claim that $T(u_0) = j$ is probabilistically relevant to $Y(u_0) = y$. All of these problems are interesting, but they are not equivalent.

[16] Dawid remarks (418) that if one assumes treatment unit additivity (but not otherwise), then causes-of-effects inferences are equivalent to retrospective effects-of-causes inferences.

The claim that $T(u_0) = j$ is probabilistically relevant to $Y(u_0) = y$ stands to the claim that $T(u_0) = j$ made such and so much difference with respect to $Y(u_0) = y$ as the ATE stands to the ITE. In any specific case, the ITE might differ from the ATE; however, estimating the ATE is still a good way to predict the value of the ITE. Similarly, a treatment might be probabilistically relevant to a response—the treatment might make a difference to the expected value of the response—and yet make no difference in the specific case under consideration. And neither being probabilistically relevant nor being a difference maker is equivalent to being the cause of an effect.

Suppose the owner of an apartment building wants to collect the insurance on it, so he attempts to set fire to the building. He does not care about the lives or property of the people in the building. Nor does he care about the risk to neighboring buildings or their residents. He soaks the walls of the basement of his building with accelerant and drops some lit cigarettes to start a fire. But before a fire starts, a tenant discovers him and prevents any catastrophe. The actions of the apartment owner raised the probability of a fire, but they did not actually cause a fire. In Pennsylvania, the apartment owner would be guilty of attempted arson (a misdemeanor) but not of arson (a felony carrying a more serious penalty). And Pennsylvania is not unusual in making such a distinction.

1.4.4 The Problems of Pre-Emption and Over-Determination

In the previous section, I claimed that Dawid conflates simple difference-making with actual causation. In this section, I want to illustrate the difference between actual causation and simple difference-making by discussing the problems of pre-emption and over-determination. Whereas, a difference maker is always an actual cause of some effect, not every actual cause of an effect is

a difference maker. Cases of pre-emption and over-determination are examples of causation without difference-making.

First, consider a case of pre-emption. Suppose that two assassins, Ralph and Lauren, are hired to kill King Victim. Ralph is supposed to poison Victim, but just in case he fails, Lauren has rigged an explosive device with a remote detonator. As it happens, Ralph poisons Victim's wine, and while Lauren watches from the shadows, the king dies. Had King Victim not died from the poisoning, Lauren would have triggered the explosive, which certainly would have killed the king. Now, after the king dies, the medical examination reveals that he died from poisoning, and investigators manage to put together an account of the mechanism by which the poison was administered. So, it seems that Ralph and the poison caused the king to die.

However, if Ralph had not poisoned the king, then Lauren would have triggered the explosive, and the king would have died all the same. Model Ralph and Lauren as treatment variables and the king as a response variable. For both treatments, the ITE is equal to zero. Hence, neither counts as a cause of the king's death according to a simple difference-making theory of actual causation. For that reason, the simple difference-making account of cause-of-effect reasoning appears to be inadequate. In some cases, we can have causation without difference-making.

Second, consider a case of over-determination. This time, suppose that Ralph and Lauren decide to kill King Victim by shooting him with a rifle from a long way off. They both fire at the same time and both of their shots strike Victim in the heart at the same moment. Assume that either shot would have killed the king on its own, which is not too much of a stretch for a bullet to the heart. The king's death is over-determined by the actions of the two assassins. Again, for both treatments the ITE is zero. But we do not want to say that the king's death was uncaused!

No, both actions appear to be actual causes of the king's death. So, again we have an example of causation without difference-making.

1.5 SHAFER'S CRITICISMS: SEGUING TO ACTUAL CAUSATION

Shafer (2000) generally praises Dawid for his clarity and skeptical attitude towards counterfactuals and other metaphysical elements in statistical practice. However, he does not think that Dawid goes far enough, especially with respect to inferences about the causes of effects. And he raises a number of related objections to Dawid's approach to reasoning about the causes of effects. Shafer writes:

> [Dawid] concedes too much by agreeing to pose a question that has no meaning. Dawid poses the general question in these words: "We are interested in whether, for the specific unit u_0, the application of *t* 'caused' the observed response." He lets us know, with the quotation marks around *caused*, that he is asking a silly question. Unfortunately, the quotation marks do not save him from becoming entangled in silly answers. … In practice, I am willing to trust Phil Dawid to take the air out of silly questions by showing how they depend on the arbitrariness of a context. But I distrust his formulation, for it seems to say that all singular causal questions partake in this silliness—that all causal answers depend on the arbitrary specification of concomitants. (440-441)

In Dawid's defense, attorneys, historians, diagnosticians, and others have been posing cause-of-effect questions for a long time. Dawid is not posing a novel question; he is attempting to say what might constrain an answer to a very old question. That the solution to an arbitrary causes-of-effects problem depends on the model one chooses (or in Dawid's case, on the concomitants one chooses) is neither surprising nor unsatisfying. Some models will do a better job of answering our questions than will other models.

Shafer claims that reasoning about the causes of effects is arbitrary, though his examples are not especially helpful. For example, he writes:

> Imagine that there was a categorical rule about the effect of aspirin on headaches:
> - At least two aspirin with at least a cup of water: the headache goes away.

- Less aspirin or less water: the headache persists.

I take the requisite aspirin with the requisite water. My headache goes away. Is it because I took aspirin? I understand the causal structure perfectly, but cannot answer the question with a simple yes or no. (441)

Shafer's puzzlement is itself puzzling. In the imagined example, the aspirin is a difference maker. Dawid's formulation of the problem applies without any modifications. Given that we know the details of the causal structure, the answer seems perfectly obvious. Yes, the headache went away because of the aspirin. The aspirin-taking caused (or was a cause of) the headache-leaving.[17]

However, Shafer might be flagging another ambiguity in causes-of-effects questions. Lewis (1973b) notes that in ordinary conversation, we sometimes pick out a single event and call it *the* cause, while relegating other contributing causes to the background and calling them mere conditions. Maybe Shafer is calling attention to the fact that in cases where more than one thing has to operate at the same time in order to produce an effect, neither one by itself should count as *the* cause: at least, not without reservations or hedges. Along these lines, Shafer writes: "It is equally silly to isolate a single action and ask whether it is *the* cause when the action's effect depends on something that is settled later—what Dawid calls a 'determining concomitant'" (441).

Shafer's most interesting criticism, however, comes back to his accusation that Dawid makes all singular causal claims arbitrary. Later in his critical piece, Shafer goes on to raise doubts about the very meaningfulness of causes-of-effects questions. He writes:

[17] One might worry about the possibility that the aspirin is pre-empted by something else. Maybe a mad scientist shoots my brain with a cosmic ray gun that makes my headache go away just as I begin to swallow the aspirin (but before it has had any time to actually work). Such a story is possible (well, maybe not with a "cosmic ray gun"), but it is not consistent with the details of the causal structure described by Shafer. Provided Shafer has correctly described the causal structure, no pre-emption occurs, and the aspirin-taking caused the headache-leaving. What counts as a cause of some effect is relative to a model. But this kind of model relativity is not vicious.

> [Consider] a very simple example. I am required to bet $1 on the outcome of a toss of a coin. I decide to bet on heads, the coin lands tails, and so I lose my $1. Did my choice of heads "cause" my handing over $1 instead of receiving $1?
>
> Here again, the causal structure is perfectly understood. The coin is fair; the chance of its landing heads is 50% regardless of how I bet. The outcome *Y* (which will be either +1 or –1) is completely determined by the treatment *T* (which will be either "bet on heads" or "bet on tails") together with the determining concomitant *D* (which will be either "coin lands heads" or "coin lands tails"). I understand exactly what happened. But this does not enable me to give a yes or no answer to the question whether *T* "caused" *Y*.

One might think that Shafer is just confused about whether or not the counterfactuals at stake are backtracking. Once the counterfactuals are clearly understood as *non*-backtracking, the answer to the question, "Did my choice of heads 'cause' my handing over $1 instead of receiving $1?" seems clear. Yes, it did. Only if the counterfactual is backtracking could the answer be, "No."

But Shafer pushes on and raises two more interesting points. Facts about causal structure *constrain* but do not generally *determine* what the causes of effects are. In the cases Shafer describes, we know what the causal structure looks like (by hypothesis). As a consequence, we know what would happen as the result of various manipulations. Hence, we can make predictions about proposed interventions.[18] What more would be added to our knowledge by identifying some occurrence as a cause or even as *the* cause of a given effect? What purpose would it serve to identify a cause of some effect? Suppose God told us that Shafer lost because he bet on heads. What practical value would that information have for us?

Shafer's second point is related to Goodman's new riddle of induction and the problem of projectable predicates (Goodman 1983). In Dawid's approach to causes-of-effects reasoning, whether or not some treatment counts as a cause depends on the concomitants one chooses. Different choices lead to different causal judgments. In order to make his point, Shafer switches

[18] We cannot make retrodictions about counterfactual interventions without falling into the kind of model-dependence that Shafer complains about, but if we are willing to trust a model, then we can make such retrodictions.

to an alternative causal structure in which the bet not only causes the eventual payout but also causes the coin to come up the opposite way of however he bet. He writes:

> Had I instead bet on tails, would the toss have come out the same way? What is gained by asking this question or by making up an answer to it? One can make up whatever answer one wants. If one assumes that this particular determining concomitant *D* comes out the same in the counterfactual world where I bet on tails, then I win in that world. If I choose a different determining concomitant D′, whose possible values are "coin lands the way I bet" and "coin lands opposite the way I bet," and assume that it comes out the same in the counterfactual world, then I lose in that world. What is the point or content of either assumption? (441)

Judgments about the causes of effects depend on model specification. However, when concomitants are understood as parts of a more or less complete causal structure, it turns out that we do not have indefinite freedom in which variables we pick or how they relate to one another.

The right solution, I think, is to understand the place of concomitant variables in causal structure and the relationship of causal structure to judgments about the (actual) causes of effects. As we have seen from Dawid, Pearl, and Wasserman, there are several different kinds of causes-of-effects reasoning problem: identifying the difference-makers relative to a given effect, estimating the amount that a potential cause raises the probability of a given effect, and identifying the actual cause(s) of a given effect. I claim that the most important cause-of-effect problem—the problem that both statisticians and metaphysicians are implicitly trying to solve—is the problem of identifying the actual cause(s) of a given effect, which I will call simply *the problem of actual causation*. In Chapter 2, I take a baby step toward a solution to the problem of actual causation by distinguishing between structural causation and actual causation.

CHAPTER TWO

STRUCTURAL CAUSATION AND ACTUAL CAUSATION

In Chapter 1, I introduced three varieties of causal reasoning problem, and I reviewed Dawid's (statistical) approach to reasoning about the causes of effects. I argued that Dawid and his interlocutors use "cause of effect" in significantly different ways, and I claimed that the most interesting causes-of-effects problem—the problem of actual causation—is best articulated in the language of graphical causal modeling. In this chapter, I look at the three varieties of causal reasoning problem from a structural-equations perspective. I distinguish structural causation from actual causation, and I relate the different varieties of causation to some other important concepts in the causation literature, including the distinction between singular (or token) causation and generic (or type) causation.

Here is how I will proceed. In Section 2.1, I describe the mathematical machinery of structural equations and the relationship of that mathematical machinery to causal structure in the world. In Section 2.2, I give an overview account of actual causation. I will set out and critique specific theories of actual causation in Chapter 3. In Section 2.3, I discuss causal generalizations. In Section 2.4, I remark on the problem of mixed populations. I point out that causal structure constrains actual causal relations, and I make a small contribution to graphical notation.

2.1 STRUCTURAL EQUATIONS AND CAUSAL STRUCTURE

Let U, called *the universe* (*of discourse*) or *population*, denote an arbitrary set of *units*, u_i. Units might be people, states, universities, actions, events, processes, or anything else one might be interested in. A *random variable* (or simply a *variable*) is a measurable function from U into the real numbers. Random variables typically represent properties of units, and the value of a variable X for u_i, denoted $X(u_i) = x$, represents the result of a measurement of the property represented by X taken with respect to the unit u_i. For example, a random variable might represent height in meters or annual operating budget in dollars. A unit may be regarded as having (or being) a collection of measurable properties. Whenever a variable takes a unit as its argument, the variable indicates which property value the unit has. For example, let U be the set of residents of the city of Riverside, and let John Smith be a resident of Riverside. Let H be a random variable representing height in meters, and let $\$$ be a random variable representing annual income in dollars. Suppose that H(John Smith) = 1.76 and $\$$(John Smith) = 30,000. Hence, John Smith is represented as having the properties "being 1.76 meters tall" and "having an annual income of $30,000."

A *structural equation model* (SEM) is a collection of equations in which (1) the independent variables in a given equation are interpreted as causes of the dependent variable in that equation and (2) the dependent variable in one (or more) of the equations may appear as an independent variable in one or more of the equations in the model.[19] Equations in an SEM contain random variables and parameters, called *structural parameters*.[20] The random variables in an SEM are divided into *error terms* and *substantive variables*.[21]

[19] See Bollen (1989) and Kline (1998) for introductions to SEMs. I have (mostly) suppressed the statistical details in my treatment of SEMs, since the problems I am engaged with here do not involve probabilities.
[20] Some SEMs contain non-random variables as well, but we will not consider such models.
[21] The term "substantive variable" is due to Spirtes.

Let $\mathcal{M} = \langle \boldsymbol{V}, \mathcal{F} \rangle$ denote an arbitrary SEM, where $\boldsymbol{V}$ is a vector of random variables and $\mathcal{F}$ is a set of structural equations involving the variables in $\boldsymbol{V}$. All of the variables in a given structural equation model map U into the real numbers. In other words, every SEM is a model relative to some universe of discourse. In ideal conditions, a structural equation model is an abstract representation of a *causal system*, which is a part of the world and involves real-world properties along with the causal relations between (or among) those properties.[22] In what follows, I will use the term *causal structure* to refer ambiguously to the structural properties of a causal model and also to the structural properties of a causal system. If an SEM $\mathcal{M}$ correctly represents the causal system for some unit u, then the unit u is said to satisfy the model $\mathcal{M}$. If each unit in U satisfies the model $\mathcal{M}$, then U satisfies $\mathcal{M}$.[23] The set U might be large or it might be small. In fact, U might be a singleton set. Whenever U is a singleton set, call the SEM over U a *singular* model. A singular SEM is deterministic in the following senses: (1) each variable in the model takes a determinate value, (2) the observed values of the substantive variables are fixed given the values of the error terms, and (3) the model exactly predicts the results of interventions on the system.[24]

Going forward, I will pretend that all of the SEMs being considered support a causal interpretation and hence count as *causal models*. In other words, the equations in an SEM are treated as what Freedman (2005, 85-95) calls *response schedules*. An equation is a response schedule if it predicts the values that the dependent variable would take if the independent

[22] In practice, the fact that one has fit an SEM to some data does not mean that the resulting model succeeds in representing a causal system. The model might not support a causal interpretation for various reasons, e.g. the data might not have been generated from a single process. In such cases, an SEM is just a very abstract kind of statistical model.

[23] The formulation I am giving here supposes that the units are homogeneous with respect to the models. However, model satisfaction might be defined in such a way as to allow heterogeneity. If the units are not homogeneous with respect to the model, then the model is satisfied in distribution.

[24] Some generalizations of SEMs do not assume determinism. See Spirtes (1995).

variables in the equation were set to some specified values. In other words, the parameters in a response schedule are invariant under interventions on the independent variables in the equation. Hence, a *structural causal relation* specifies both actual and counterfactual relations between variables for the units in a population. As the name suggests, a structural equation model represents a collection of structural causal relations.

Another way of thinking about what a structural equation model represents is in terms of idealized experiments. An experiment is a kind of intervention on a causal system. In an experiment, some properties of a unit are set to specific values and the results of that manipulation are observed. Let the manipulation of a variable X to the value x for the unit u be denoted $do(X(\mathrm{u}) = x)$. The result of manipulating a variable in an SEM is determined by replacing the equation in which the variable appears as a dependent variable with a new equation that makes the variable equal to a constant and then propagating that change through all the equations in which the variable appears as an independent variable. For example, the variable for John's annual income might have a corresponding structural equation like:

$$(2.1) \quad \$ = \beta_0 + \beta_1 \cdot \text{Education} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Gender} + \beta_4 \cdot \text{Region} + \varepsilon_{\$}$$

When John's annual income is manipulated, the equation (2.1) is replaced by simply writing down the desired value. For example, if we want to set John's annual income to $45,000, we would replace equation (2.1) with:

$$(2.2) \quad \$ = 45{,}000$$

An SEM, then, represents the results of a collection of possible experiments, and thereby, a collection of counterfactuals. Following Holland (1986) and Pearl (2000), let $Y_{X=x}(\mathrm{u})$ denote the value Y would have, for unit u, were one to manipulate the variable X to the value x with respect to unit u, i.e. if one were to $do(X(\mathrm{u}) = x)$.

Putting it all together, suppose we have the structural equation model (S1) below, characterizing the relationships among the involvement parents have with their high-school-age children (P), the number of times per week their children eat breakfast (B), and the children's SAT scores (S):[25]

$$\text{(S1)} \qquad \begin{aligned} P &= \varepsilon_P \\ B &= \beta_{B1} \cdot P + \varepsilon_B \\ S &= \beta_{S1} \cdot P + \varepsilon_S \end{aligned}$$

In (S1), the variables P, B, and S are substantive; ε_P, ε_B, and ε_S are error terms; and β_{B1} and β_{S1} are structural parameters. A model with specified values for its structural parameters is called *parameterized*; otherwise, it is called *unparameterized*. Suppose that the error variable ε_P takes values in the interval [-3, 3], the variable ε_B takes values in the interval [0, 7], and the variable ε_S takes values in the interval [600, 2400]. Let the model (S1) be a singular model over the set U_1 = {Suzy}. Suppose then that ε_P(Suzy) = 2, ε_B(Suzy) = 4, and ε_S(Suzy) = 1700. So, Suzy has interested parents, she eats breakfast $2 \cdot \beta_{B1} + 4$ times each week, and she scores $2 \cdot \beta_{S1} + 1700$ on the SAT.

Structural equation models may be represented graphically by associating each vertex in a graph with the structural equation in which the variable identified with that vertex is the dependent variable. In the graphical representation of an SEM, the error terms are typically suppressed. A variable V_p is a parent of the variable V_c in the graph $\boldsymbol{G}$ iff V_p is an independent variable in the structural equation associated with V_c. The edges in a graphical representation of an SEM are often labeled with the structural parameters in the model or with estimates of the structural parameters in the model. For example, the SEM (S1) is represented by the graph in Figure 2.1.

[25] The model (S1) is pure fiction and is being introduced for illustrative purposes only.

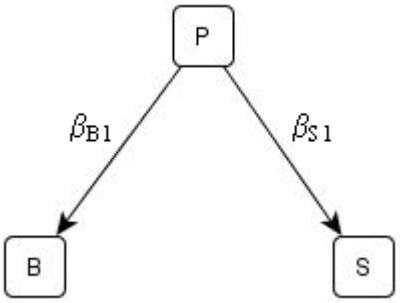

Figure 2.1: Graph for Model S1

Graphs may be given either a *statistical interpretation* or a *causal interpretation*. Under a statistical interpretation, graphical properties entail statistical properties via a Markov condition. Under a causal interpretation, graphical properties also entail facts about causal relations between or among the variables in the graph. For example, the directed edge $P \rightarrow B$ in the graph of Figure 1 entails that *P* is a direct cause of *B*. A graph under a causal interpretation is called a *causal graph*. Since we are assuming that SEMs support a causal interpretation, their graphical representations are causal graphs. As with SEMs, if a graph ***G*** correctly represents the causal system for some unit u, then the unit u is said to satisfy the graph ***G***. If each unit in U satisfies the graph ***G***, then U satisfies ***G***. A structural equation model along with its graphical representation is sometimes called a *graphical causal model*. If the parameters in a graphical causal model are given specific values, then it is called a *parameterized graphical causal model*.

The causal content of the model (S1) includes the claim that intervening to make Suzy eat breakfast five times each week would not change her SAT score, but intervening to make her parents one unit more involved (on our 0-7 scale) would raise her SAT score by β_{B1} points. Now, suppose that for Suzy, $\beta_{B1} = 1$ and $\beta_{S1} = 100$.[26] If we intervene to set parental involvement

[26] If we wanted to make the models more general, we could write the equations with the parameters as functions of the units, like β(u).

to -2 for Suzy in the parameterized version of (S1), the result of the intervention is given in the model:

$$
\begin{aligned}
&P(\text{Suzy}) = \varepsilon_P(\text{Suzy}) = -2 \\
&B(\text{Suzy}) = \beta_{B1} \cdot P(\text{Suzy}) + \varepsilon_B(\text{Suzy}) = -2 \cdot \beta_{B1} + 4 = 2 \\
&S(\text{Suzy}) = \beta_{S1} \cdot P(\text{Suzy}) + \varepsilon_S(\text{Suzy}) = -2 \cdot \beta_{S1} + 1700 = 1500
\end{aligned} \qquad \text{(S2)}
$$

As with un-parameterized SEMs, if a parameterized SEM $\boldsymbol{\mathcal{M}}_P$ correctly represents the causal system for some unit u, then the unit u is said to satisfy the model $\boldsymbol{\mathcal{M}}_P$. If each unit in U satisfies the model $\boldsymbol{\mathcal{M}}_P$, then U satisfies $\boldsymbol{\mathcal{M}}_P$.

Typically, in social and medical science applications, SEMs are *generic* models over more or less vaguely defined populations.[27] For example, an education researcher might be interested in the relationships among parental involvement, eating breakfast, and performance on the SAT among students in the United States. The researcher is not aiming to say what the causal relationships are like for Suzy but to say what they are like for the population of students in the United States. The reason is that the researcher is not especially concerned with the consequences of proposed interventions with respect to any single individual but with respect to the population as a whole. As with singular SEMs, generic SEMs are deterministic in the sense that for each unit, the variables in the model take determinate values, the observed values of the substantive variables are fixed given the values of the error terms, and the model exactly predicts the results of interventions on the system. However, different units often take different values for the same variable(s); hence, probability distributions may be defined for the variables, and the models apply probabilistically to unobserved units. Just as a causal model may be singular or generic, the causal system represented by a causal model may be singular or generic.

[27] See Freedman (2010), Chapter 2, for a discussion of the vagaries of "population" in social science work and the problems raised by those vagaries.

When talking about the causal structure of a model over a set of variables with respect to some universe U, one might have in mind: (1) a causal graph shared by all of the units in U, (2) a structural equation model for U, or (3) a parameterized structural equation model for U. Call the three kinds of causal structure graphical, functional, and parameterized, respectively. The three kinds of causal structure are hierarchically related in increasing order of strength. If U satisfies a parameterized structural equation model, then U also satisfies a corresponding un-parameterized structural equation model and a corresponding causal graph. However, U might satisfy a causal graph without satisfying a structural equation model, and U might satisfy a structural equation model without satisfying a parameterized structural equation model. Let the set of units in some universe U that satisfy some parameterized structural equation $\boldsymbol{\mathcal{M}}_P$ be called P, let the set of units in U that satisfy the structural equation model corresponding to $\boldsymbol{\mathcal{M}}_P$ be called S, and let the set of units in U that satisfy the causal graph corresponding to $\boldsymbol{\mathcal{M}}_P$ be called G. Then, for every U and every parameterized structural equation model, $P \subseteq S \subseteq G \subseteq U$. Moreover, for many universes and many models, equality does not hold between any pair of these sets.

2.2 ACTUAL CAUSATION

Statisticians, computer scientists, social scientists, and philosophers working with structural equation models, or graphical models more generally, treat causation as a structural causal relation between random variables, rather than as an actual causal relation between events. However, in the last decade, graphical models have been adapted, by Pearl (2000), Hitchcock (2001a), Woodward (2003), Halpern and Pearl (2005), Glymour and Wimberly (2007), Hitchcock (2007), and Halpern (2008), to produce accounts of actual causation. These accounts

disagree about a number of specific details, but they also have some wide agreement, which I want to consider in this section.

Consider an ordered set $\boldsymbol{V}$ of variables, partitioned into the variables X and Y along with the ordered set $\boldsymbol{Z}$ of variables obtained by removing X and Y from $\boldsymbol{V}$. Say that $X(\cdot)$ is a *direct structural cause* of $Y(\cdot)$ relative to the population U and the ordered set $\boldsymbol{V}$ of variables if for each u in U, there exist values x_1 and x_2 of X and values $\boldsymbol{z}$ of $\boldsymbol{Z}$ such that $x_1 \neq x_2$ and $Y_{X=x1,\, \boldsymbol{Z}=\boldsymbol{z}}(\mathrm{u}) \neq Y_{X=x2,\, \boldsymbol{Z}=\boldsymbol{z}}(\mathrm{u})$. In other words, the variable $X(\cdot)$ is a direct structural cause of the variable $Y(\cdot)$ if there is a pair of $do(\cdot)$ operations such that the value of $Y(\mathrm{u})$ given $do(X(\mathrm{u}) = x_1, \boldsymbol{Z}(\mathrm{u}) = \boldsymbol{z})$ differs from the value of $Y(\mathrm{u})$ given $do(X(\mathrm{u}) = x_2, \boldsymbol{Z}(\mathrm{u}) = \boldsymbol{z})$. The variable X is a *structural cause* of the variable Y iff X is a direct structural cause of Y in the reduced set of variables containing only X and Y.[28] Actual causation requires something more than the existence of a pair of $do(\cdot)$ operations satisfying the definition of structural cause. However, if such a pair does not exist for variables X and Y, then there can be no actual causal connection between any values of X and any values of Y. Structural causation is a necessary precondition of actual causation.

Recall the assassins Ralph and Lauren. Suppose that U is a collection of assassinations carried out by Ralph and Lauren working in tandem. Each unit u is a single assassination. Suppose that for the assassinations in U, Ralph and Lauren take turns killing their victims. Sometimes Ralph takes the lead and Lauren acts as backup. Sometimes Lauren takes the lead and Ralph acts as backup. And suppose that for each unit in the population, whoever takes the lead is successful. Let the variable $R(\cdot)$ represent Ralph's action such that for all u, if Ralph acts, then $R(\mathrm{u}) = 1$ and if Ralph does not act, then $R(\mathrm{u}) = 0$. Similarly, let the variable $L(\cdot)$ represent

[28] One might be tempted (as I was in an earlier draft) to define structural causes in terms of chains of direct structural causes. For example, one might say that X is a structural cause of Y iff there are variables $\boldsymbol{Z} = \{Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}\}$ such that X is a direct structural cause of $Z_{(1)}$, $Z_{(i)}$ is a direct structural cause of $Z_{(i+1)}$ for $0 < i < n$, and $Z_{(n)}$ is a direct structural cause of Y. However, that definition assumes transitivity of structural causation, which might fail if a causal structure is unfaithful.

Lauren's action. Moreover, let the variable $V(\cdot)$ represent the state of the victim such that for all u, if the victim is alive, then $V(u) = 1$ and if the victim is dead, then $V(u) = 0$. For present purposes, suppose that both Ralph and Lauren are perfect assassins, so that if either one acts, the victim dies. Hence, for each u, one may write $V(u) = R(u) + L(u)$, where '+' is the Boolean OR.

The variable $R(\cdot)$ is a direct structural cause of $V(\cdot)$, since $V(u) = 1$ given do($R(u) = 1$, $L(u) = 0$) and $V(u) = 0$ given $do(R(u) = 0, L(u) = 0)$. In the same way, $L(\cdot)$ is a direct structural cause of $V(\cdot)$, since $V(u) = 1$ given do($R(u) = 0$, $L(u) = 1$) and $V(u) = 0$ given $do(R(u) = 0$, $L(u) = 0)$. Despite the fact that only one of Ralph and Lauren acts in any actual case, both assassins are structural causes of the state of their victim in each assassination. The graph corresponding to the Ralph and Lauren case is given in Figure 2.2.[29]

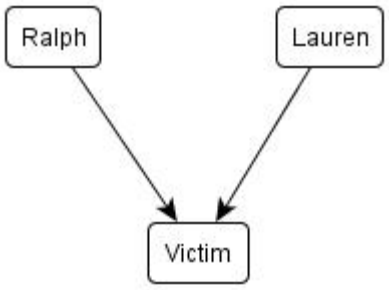

Figure 2.2: Causal Graph for the Ralph and Lauren Case

Here and elsewhere, I have drawn graphs corresponding to structural equation models. Before going any further, I want to give a bit more technical detail. Let a *directed graph* be an ordered pair $\boldsymbol{G} = \langle \boldsymbol{V}, \boldsymbol{E} \rangle$, where $\boldsymbol{V}$ is a finite set of *vertices* and $\boldsymbol{E} \subseteq (\boldsymbol{V} \times \boldsymbol{V})$ is a finite set of directed *edges*. An edge $\langle V_1, V_2 \rangle$ is *directed from* V_1 *into* V_2. Denote the directed edge $\langle V_1, V_2 \rangle$ by $V_1 \rightarrow V_2$. A *path* of length $n > 0$ from V_i to V_j, denoted $V_i \mapsto V_j$, is a sequence $V_{(1)}, V_{(2)}, \ldots, V_{(n+1)}$ of vertices such that $V_i = V_{(1)}$, $V_j = V_{(n+1)}$, and $V_{(k)} \rightarrow V_{(k+1)}$, for $k = 1, \ldots, n$. The graph in Figure

[29] One cannot tell that the case is a case of pre-emption (as opposed to over-determination) on the basis of the graph alone. That is because pre-emption and over-determination are problems that belong specifically to *actual causation*, and causal graphs represent *structural causation*, not actual causation.

2.1 is not very complicated. It has two paths of length one—one from Ralph to Victim and one from Lauren to Victim.

Now, consider a specific assassination, like the example of pre-emption in Section 1.3.4. Ralph poisons King Victim. Lauren merely watches as Victim dies; however, Lauren was prepared to detonate an explosive that would surely have killed the king in event that Ralph's poison failed or some other mischance took place. Lauren's act is pre-empted by Ralph's act.[30] Structural causation does not generally seem to capture our intuitions about the actual cause(s) of a given outcome. Whereas structural equation models treat causation as a relation between random variables, theories of actual causation specify conditions under which a random variable taking on some value causes another random variable to take on some (other) value. A theory of actual causation tells us whether $V_c(u) = v_c$ counts as an actual cause of $V_e(u) = v_e$.[31]

Actual causation puts stronger constraints on how one quantifies over $do(\cdot)$ operations with respect to some collection of variables. For a variable X to count as a structural cause of another variable Y, it is sufficient to find a pair of values x_1 and x_2 for X such that holding everything else fixed, Y takes on two different values when X takes on its different values. Actual causation puts more restrictions on the $do(\cdot)$ operator. Current theories of actual causation restrict attention to actual causation in the singular case, but quantification over $do(\cdot)$ operations is orthogonal to quantification over units. A collection of structural causation

[30] Such cases are sometimes called cases of *early* pre-emption to distinguish them from harder cases in which two or more causal processes, each of which would be sufficient to produce the actual effect, are initiated but only one runs to completion. The harder cases are called cases of *late* pre-emption. As an example of late pre-emption, suppose that Ralph and Lauren both shoot at King Victim, but Ralph shoots two milliseconds after Lauren. Lauren's bullet reaches Victim one millisecond before Ralph's bullet. Both bullets pass through Victim's heart, and both would have been sufficient to kill Victim. Hence, Lauren's shot is a late pre-emptor of Ralph's shot with respect to Victim's death.

[31] In some accounts, for example Hitchcock (2007a) and Halpern (2008), $V_c(u) = v_c$ and $V_e(u) = v_e$ are understood as singular (or token) events. Although equations of random variables pick out events in the statistical sense, I think that the semantics for random variables requires that $V_c(u) = v_c$ and $V_e(u) = v_e$ be understood as property-instances, not events.

relations might hold for a large population or for a small population (maybe even for a single unit). Similarly, a collection of actual causation relations might hold for a large population or for a small population (maybe even for a single unit).

2.3 CAUSAL GENERALIZATIONS

I have been at some pains to point out that both structural causal relations and actual causal relations might hold for large or small populations. The reason is that many philosophers describe actual causation as token or singular causation, without being very careful about what an actual causal relation is a token or instance of, exactly. The relata of an actual causal relation are evaluated random variables, so maybe by identifying actual causation with token (singular) causation and identifying structural causation with type (generic) causation, philosophers mean to say that whereas structural causal relations relate generics or types of events or some such, actual causal relations relate singulars or token events or some such. However, when philosophers talk about actual causation, they always produce examples involving a single unit. For example:

(S) Suzy had a high score on the SAT because she ate breakfast regularly.

And they contrast those examples with causal generalizations like:

(G) Eating breakfast causes higher SAT scores.

Which are generic both with respect to the units involved—the generalization in (G) applies to a vague population—and also with respect to the causal relata. These are, of course, different ways in which a causal relation might be generalized, but they are often run together. Hitchcock (2001b) notices the problem and claims that "the distinction between singular and general causation conflates two separate distinctions, and … this conflation has greatly impeded progress

in the understanding of causal generalizations" (222). But despite his own warning, in subsequent papers, he has not kept the two distinctions clearly separated.

Actual causal relations may be the subject of practically interesting generalizations. Moreover, different populations that share the same causal structure sometimes support different (probabilistic) generalizations about actual causation. For example, consider again the story of the assassins Ralph and Lauren. Suppose $R(u) = 1$ and $L(u) = 0$ for most $u \in U_1$. Then it will be true, relative to U_1, that $R(u) = r$ is probably an actual cause of $V(u) = v$ and $L(u) = l$ is probably not.[32] That is, most units in the universe U_1 will be such that Ralph is the actual cause of Victim's death. By contrast, if $R(u) = 0$ and $L(u) = 1$ for most $u \in U_2$, then it will be true, relative to U_2, that $L(u) = l$ is probably an actual cause of $V(u) = v$ and $R(u) = r$ is probably not. However, in universe U_1 and also in universe U_2, both R and L are structural causes of V.

2.4 MIXED POPULATIONS AND INTERACTIONS

Starting with a structural causal model with respect to a population, if one restricts attention to a sub-population, one obtains a (possibly different) structural causal model. A structural causal model does *not* become an actual causal model simply by restricting the population. More information is needed in order to identify the actual causes, even if the structural causal model is singular. However, restrictions on the population might make a difference to what causal structure and what actual causal relations hold. For example, Simons et al. (2009) looked at the influence of religiosity on some risky sexual behaviors among adolescents, specifically age at sexual debut, number of sexual partners, and seriousness of commitment between sex partners. They used pencil and paper surveys to collect data from 2,109 undergraduates in sociology (enrolled in the academic year 2001-2002). They measured religiosity, sexually permissive

[32] Here, "probably" is being used to express the notion that the probability of the proposition p is greater than 0.5.

attitudes, level of commitment to first sex partner, feelings about first intercourse, age at first intercourse, and number of premarital sex partners. They found that the causal structure for females (pictured in Figure 2.3) differed from the causal structure for males (pictured in Figure 2.4).

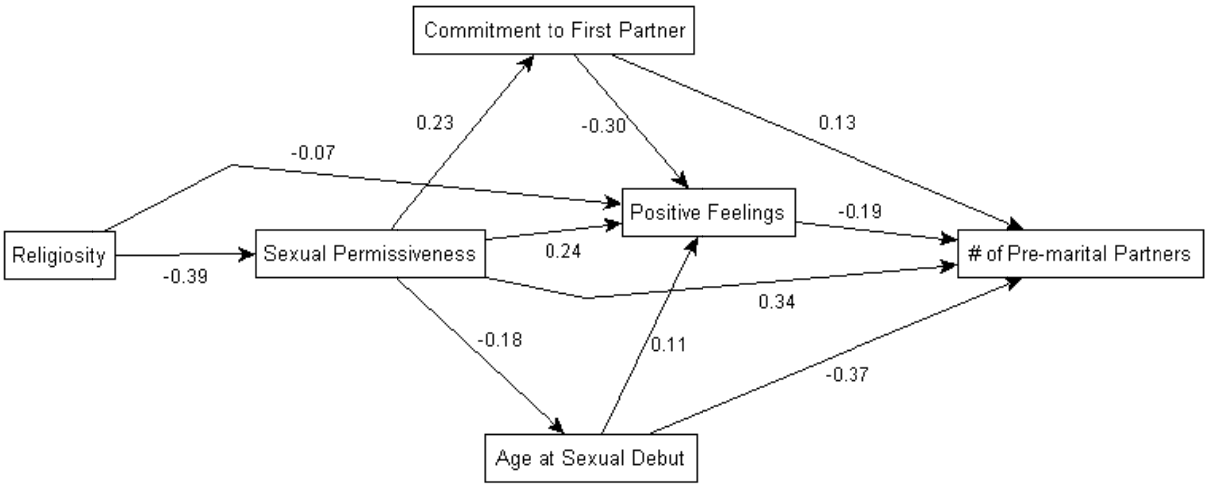

Figure 2.3: Simons et al.'s Model for Females

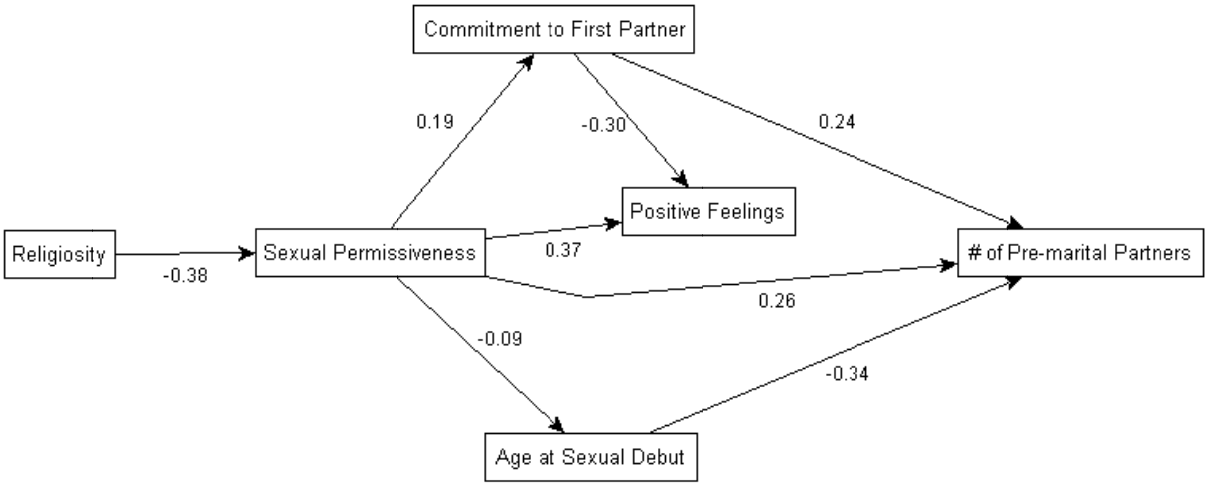

Figure 2.4: Simons et al.'s Model for Males

The models in Figures 2.3 and 2.4 are structural causal models: they relate property-universals. Partitioning the units into males and females does not yield a model of actual causation for any

specific unit or group of units. As I have noted several times, a structural causal model over a restricted population, even a singleton population, is not equivalent to an actual causal model.

However, a pair of structural causal models might tell us something interesting about how the actual causes differ in the two populations, since an evaluated variable $X = x$ cannot be an actual cause of another evaluated variable $Y = y$ (relative to U) unless X is a structural cause of Y (relative to U). In the models of Figures 2.3 and 2.4 from Simons et al., some value for Positive Feelings might be an actual cause of some value for Number of Pre-marital Partners for some females, but the value of Positive Feelings cannot be an actual cause of Number of Pre-marital Partners for any males, unless the population of males is heterogeneous.[33]

Cases like that reported by Simons et al., which are common in social, medical, and environmental sciences, may be treated in two different ways. One might treat the different structural models in the different sub-populations as reflecting heterogeneity in the population at large, as I have done above. From this perspective, it makes sense to write down different structural models for the (structurally) different sub-populations. Alternatively, one might treat the different sub-populations as different values for a variable that has been left out of a more complete model over the entire population.

But how do we recover the structurally different models over the sub-populations from a single model over the whole population? Answering that question also resolves a tension that I see in the policy-variable representation of interventions in graphical causal modeling.[34] Suppose we want to determine whether a treatment X has any effect on a response Y. And

[33] The point is not restricted to direct causes. If Number of Pre-marital Partners directed structurally caused another variable, say Trust of Present Partner, in both models, then it would be true that for females but not for males, some value for Positive Feelings might count as an actual cause of some value for Trust of Present Partner.
[34] See Spirtes et al. (2000) and Eberhardt and Scheines (2007) for presentations of the policy variable approach.

suppose the causal system has the structure in Figure 2.5, where L is an unobserved common cause of X and Y.

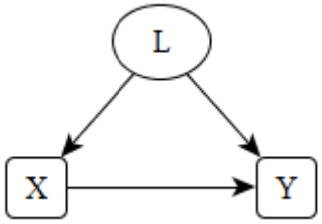

Figure 2.5: Causal Graph with Confounder

In order to exclude the possibility that an observed association between X and Y is due to the confounder L, we conduct an experiment in which we assign units to treatment and control groups, which are assigned different values of X. In the policy-variable approach, this scenario is represented by adding a *policy variable* that is a structural cause of the variable X but of no other variable in the graph. When the policy variable has the value "1," all of the edges directed into X are cut out of the causal graph, as in Figure 2.6.

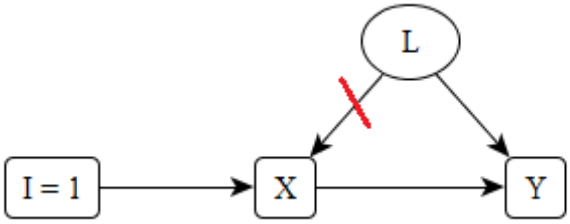

Figure 2.6: Causal Graph with Policy Variable Intervention

In the case being considered, the resulting graph in Figure 2.7 has one fewer edge than the original graph in Figure 2.5.

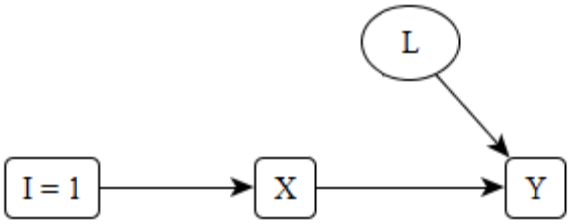

Figure 2.7: Post-Intervention Graph

On the other hand, when the policy variable has the value "0," the original graph is left just as it was, except that it is augmented with the policy variable, as in Figure 2.8.

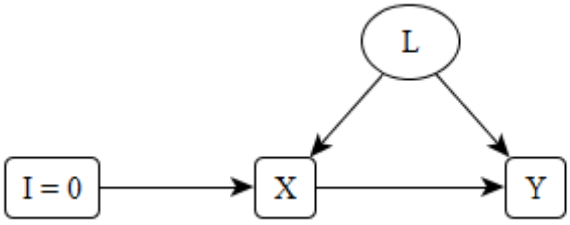

Figure 2.8: No-Intervention Graph

The policy-variable approach to interventions is easy to understand and apply. But what is the formal mechanism by which the edge from *L* to *X* disappears in the case $I = 1$ but not in the case $I = 0$? The latent variable should be a structural cause (or not) regardless of the value that any other variable takes on in the graph. Whether it is an *actual* cause might depend on the specific values that other variables take, but whether it is a structural cause should not.

I propose that we think about interventions in a slightly different way, which connects them to interactions and makes sense of graphs over mixed populations, like the population of college students in Simons et al. (2009).[35] Redefine the vertices of a graph to be a finite subset of the cross-product of the power set of the set of vertices, $\boldsymbol{E} \subseteq \mathscr{P}(\boldsymbol{V}) \times \mathscr{P}(\boldsymbol{V})$, where $\mathscr{P}(\boldsymbol{V})$ is the

[35] See Frey (2003) and Pearl and Bareinboim (2010) also develop alternative graphical notations. Pearl and Bareinboim's "switch graphs" are closer to mine in function; Frey's "directed factor graphs" are closer to mine in form.

power set of $\mathbf{V}$. When $\mathbf{V}_1 = \{V_1\}$ and $\mathbf{V}_2 = \{V_2\}$ are singletons, call $<\mathbf{V}_1, \mathbf{V}_2>$ a *main-effect edge*, and write $<V_1, V_2>$. Let $<\{V_1, V_2, \ldots, V_n\}, \{V_{n+1}\}>$ denote the *interaction edge* from the vertices $V_1, V_2, \ldots, V_n$ into the vertex V_{n+1}.

An interaction edge from $V_1, V_2, \ldots, V_n$ into V_{n+1} represents an interaction term $V_1V_2\ldots V_n$ in the equation for V_{n+1}. As a simple illustration, consider the equation $Y = \mathrm{a}\cdot X_1X_2 + \varepsilon$. The corresponding graph in Figure 2.9 has a single interaction edge $<\{X_1, X_2\}, Y>$.

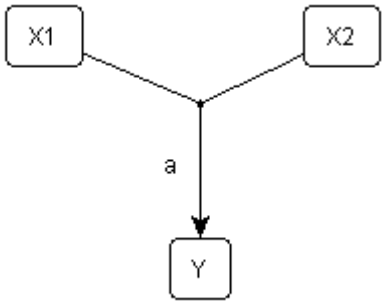

Figure 2.9: Simple Interaction Graph

Typically, each equation in a structural equation model has main-effects terms. The equation $Y = \mathrm{a}\cdot X_1 + \mathrm{b}\cdot X_1X_2 + \mathrm{c}\cdot X_2 + \varepsilon$ corresponds to the graph in Figure 2.10.

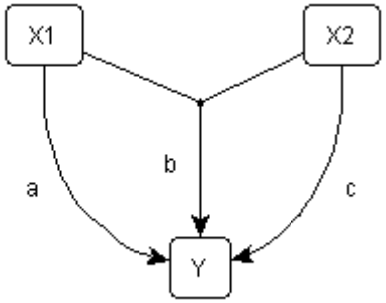

Figure 2.10: Interaction Graph with Main Effects Edges

I am now in a position to reformalize interventions. Instead of a single policy variable, introduce two variables, *I* and *P*, where *I* represents the bare fact that there is or is not an intervention, and

P represents the details of the policy. Suppose that we have the simple causal structure in Figure 2.11.

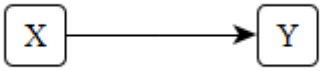

Figure 2.11: A Simple Causal Graph

Now, suppose we want to model an intervention on *Y*. We write $Y = \alpha (1 - I) X + I \cdot P$, and we draw the corresponding intervention graph in Figure 2.12.

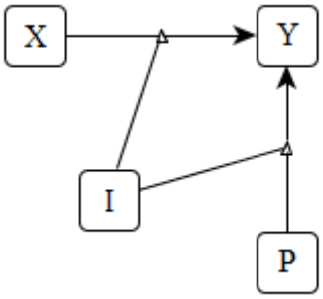

Figure 2.12: Illustration of the Interaction Model for Interventions

The variable *I* is allowed to take the value 0 or 1; the variable *P* is allowed to take any value that *Y* is allowed to take. If we want to set *Y* to the value *y*, we let $I = 1$ and $P = y$. Now, we add a general rule for interpreting interaction edges: When a variable that is part of an interaction term takes on the value 0 for every unit in U, remove the interaction edge from the graph

2.5 SOMETHING MORE

Actual causation is not simply structural causation restricted to a singleton population. Structural causation and actual causation are distinct but related modes of causal relation. Structural causation is a necessary pre-condition for actual causation: $X = x$ cannot be an actual cause of $Y = y$ unless *X* is a structural cause of *Y*. Actual causation places constraints on the $do(\cdot)$ operator

that are strictly stronger than the constraints placed on the *do*(·) operator by structural causation. In short, actual causation is structural causation plus something. Exactly what the "something" is remains unclear. The current state of the art in actual causation research is looking for the something more.[36] In the next chapter, I motivate the continued search for something more by showing that a range of current theories have not yet found it.

[36] See Hall (2007), Hitchcock (2007a), Halpern (2008), and Hitchcock and Knobe (forthcoming).

CHAPTER THREE

ACTUAL CAUSATION AND SIMPLE VOTING SCENARIOS

In Chapter 2, I discussed the difference between structural causation and actual causation, but I did not provide any detailed theories of actual causation. In the present chapter, I describe a range of current theories of actual causation, and I produce counter-examples to those theories. The counter-examples come from applying the theories of actual causation to some quite ordinary voting scenarios and note that the theories say rather strange things about them. Two prima facie examples of defects in the theories are these: (1) in all simple-majority elections that allow abstentions, the theories of actual causation under consideration count every abstention as an actual cause of the winning candidate's victory, regardless of whether the election is closely contested; and (2) in all simple-plurality elections involving three or more candidates, every theory of actual causation under consideration counts every vote as an actual cause of the winning candidate's victory, regardless of how the votes are actually distributed among the candidates.

The complications introduced by the discussion in this chapter are only the latest in a long sequence. Already, the theories of actual causation under consideration have been outfitted against well-known counter-examples. In fact, the theories say the strange things they do about voting due to an over-reaction to the problem of over-determination. The theories of actual causation under consideration accommodate the *individualist intuition* that when some outcome is over-determined by two or more occurrences, each occurrence is a cause of the outcome. In

order to accommodate the individualist intuition, the theories have to take account of facts about difference-making in the actual circumstances *and* in counterfactual scenarios. However, the theories are too permissive about the range of counterfactual scenarios they consider. The theories do not have the resources to block enough counterfactual scenarios from consideration (or at least they do not have the resources to block the right ones).

Here is how I proceed. In Section 3.1, I describe three different theories of actual causation from the contemporary literature. In Section 3.2, I apply these theories to some simple voting scenarios. In Section 3.3, I argue that the deliverances of the theories with respect to voting scenarios are counter-intuitive. Hence, simple voting scenarios offer prima facie counter-examples to current accounts of actual causation.

3.1 THEORIES OF ACTUAL CAUSATION

In this section, I will describe three distinct attempts to provide necessary and sufficient conditions for determining whether a given variable taking on some specific value for a specific unit is an actual cause of another variable taking on some (other) specific value for that unit.

3.1.1 Hitchcock and Woodward

Hitchcock (2001) and Woodward (2003) have developed very similar theories of actual causation. Though they are not equivalent in general, they are equivalent with respect to the examples I develop in Section 3 below. Thus, I provide a single statement representative of their theories with respect to my examples. Begin with the simple theory that $X(\mathrm{u}) = x$ is an actual cause of $Y(\mathrm{u}) = y$ iff the following two conditions are satisfied:[37]

[37] See Woodward (2003, 74-77) and Hitchcock (2001, 286-287) for their descriptions of the simple theory.

(HW1) The actual value of X is x, and the actual value of Y is y, for unit u.

(HW2) There exists a path P from X to Y and there exists a manipulation $do(X = x^*)$ for some $x^* \neq x$ such that $Y_{X=x^*}(u) \neq y$ whenever all variables not on the path P are held fixed at their actual values.

Unfortunately, (HW2) is not satisfied if the value of Y (the putative effect) is over-determined by independent causal mechanisms. Thus, there are no actual causes of over-determined events according to the simple theory proposed by Hitchcock and Woodward. Although some philosophers, notably Lewis (1986, Appendix E), have been willing to accept this consequence, Lewis' intuition is not very widely shared among theoreticians today. How widespread it is among ordinary speakers of English is less clear.[38] Hitchcock and Woodward both find it unsatisfactory. In order to formulate a replacement for (HW2), we need a bit more notation and another definition (due to Hitchcock). Let $\boldsymbol{w}$ denote an ordered n-tuple of values of the ordered n-tuple $\boldsymbol{W}$ of variables, and let $do(\boldsymbol{W} = \boldsymbol{w})$ denote the ordered collection of manipulations $do(W_1 = w_1), \ldots, do(W_n = w_n)$. Say that the ordered n-tuple $\boldsymbol{w}$ of values of the ordered n-tuple $\boldsymbol{W}$ of variables is in the *redundancy range* of the path P if carrying out the manipulations $do(\boldsymbol{W} = \boldsymbol{w})$ leaves all of the variables on P at their actual values. Now, in order to handle cases of over-determination, replace (HW2) with[39]

(HW2*) There exists a path P from X to Y and there exist manipulations $do(X = x^*)$ for $x^* \neq x$ and $do(\boldsymbol{W} = \boldsymbol{w})$ for $\boldsymbol{w}$ in the redundancy range of P such that $Y_{X=x^*}(u) \neq y$ whenever all the variables in $\boldsymbol{W}$ are fixed by the manipulation $do(\boldsymbol{W} = \boldsymbol{w})$.

In other words, $X(u) = x$ is an actual cause of $Y(u) = y$ if one can find some path P from X to Y and some choice of (possibly non-actual) values for all of the variables not on path P such that the variables on P retain their actual values but also such that some change in the value of X

[38] McDermott (1995) reports strong agreement with individualist intuitions among naïve undergraduates. Recent experiments that I have conducted with Justin Sytsma suggest that the picture is not so clear.

[39] See Woodward (2003, 83-84) and Hitchcock (2001, 289-290) for their descriptions of the amended theory.

would result in a change in the value of Y, if one were to set the variables not on path P to those values.

In order to check whether $X(u) = x$ is an actual cause of $Y(u) = y$, apply the following algorithm. First, pick some path P from X to Y. Second, pick some values for all the variables not on the path P such that the variables on the path P keep their actual values. (That is what it means for the off-path values to be in the redundancy range of P.) Third, set X to each of its alternative values in turn, checking whether any such change requires a downstream change in the value of Y. If any such change in the value of X results in a change in the value of Y, then stop, $X(u) = x$ is an actual cause of $Y(u) = y$. If no change in the value of X results in a change in the value of Y, then repeat the second and third steps above. Do this until all possible values for the off-path variables in the redundancy range of P have been tried. If at any stage changing the value of X results in a change in the value of Y, stop: $X(u) = x$ is an actual cause of $Y(u) = y$. Otherwise, repeat the above steps with a new path from X to Y. If no untried paths from X to Y exist, then declare that $X(u) = x$ is not an actual cause of $Y(u) = y$.

3.1.2 Halpern and Pearl

Halpern and Pearl (2005) offer a more complicated theory of actual causation. Not all of the complication matters for my examples; consequently, my presentation of their theory removes some excess. Halpern and Pearl produce two different definitions of "actual cause"; however, as was the case with Woodward's theory and Hitchcock's theory, the two definitions are equivalent with respect to my examples. According to Halpern and Pearl, $X(u) = x$ is an actual cause of $Y(u) = y$ iff the following three conditions are satisfied:

(HP1) The actual value of X is x, and the actual value of Y is y, for unit u.

(HP2) There exists a path P from X to Y and there exist manipulations $do(X = x^*)$ for $x^* \neq x$ and $do(\boldsymbol{W} = \boldsymbol{w})$ for the variables in $\boldsymbol{W}$ (all those variables not on

path P) such that $Y_{X=x^*}(u) \neq y$ whenever all the variables in $\boldsymbol{W}$ are fixed by the manipulation $do(\boldsymbol{W} = \boldsymbol{w})$.[40]

(HP3) Let $\boldsymbol{W}^*$ be an m-tuple, m ≤ n, formed from $\boldsymbol{W}$ by selecting m components of the $\boldsymbol{W}$ n-tuple, and let $\boldsymbol{w}^*$ be the m-tuple of values formed by selecting similarly indexed components of $\boldsymbol{w}$. For all possible $\boldsymbol{W}^*$, $Y_{X=x}(u) = y$ whenever all the variables in $\boldsymbol{W}^*$ are fixed by the manipulation $do(\boldsymbol{W}^* = \boldsymbol{w}^*)$.[41]

Condition (HP1) is the same as the first condition for Hitchcock and Woodward. Condition (HP2) allows for actual causation in cases of over-determination, and condition (HP3) guarantees that the assignment of values to off-path variables is not completely responsible for the change in the value of the putative effect *Y*—the change in the value of *Y* is due at least in part to the change in the value of *X*.

In order to check whether $X(u) = x$ is an actual cause of $Y(u) = y$ under Halpern and Pearl's theory of actual causation, apply the following algorithm. Pick a path from *X* to *Y*. Set the variables not on the path P to (potentially) new values such that *Y* retains its actual value *y* and so that there is some value $x^* \neq x$ such that if we set $X = x^*$, *Y* will take on a new value not equal to *y*. If such a set of values exists, then $X(u) = x$ is an actual cause of $Y(u) = y$. If there are no such values for the off-path variables, pick a new path from *X* to *Y* and try again. If all possible paths from *X* to *Y* have been tried, then $X(u) = x$ is not an actual cause of $Y(u) = y$.

3.1.3 Hall

Hall (2007) criticizes structural equation theories of actual causation on the grounds that they do not make use of any intrinsic properties of causes or effects in determining what is an actual cause of what. The intrinsic property Hall thinks we should care about is the property of being a *default* (as opposed to a *deviant*) state. For Hall, the distinction between a deviant state and a

[40] Some or all of the manipulated values of the variables in $\boldsymbol{W}$ may be the actual values of the variables in $\boldsymbol{W}$.
[41] Since Halpern and Pearl's definitions allow that an arbitrary vector $\boldsymbol{X}$ of variables may be an actual cause, they include a condition—minimality—that ensures that actual causes are as small as they can be. I have not included the minimality condition here because I will only be concerned with singleton sets in my criticisms in Chapter 4.

default state maps pretty closely onto the distinction between an occurrence (deviant state) and an absence or non-occurrence (default state). With the notion of default state in hand, Hall provides necessary and sufficient conditions for an event to count as an actual cause of another event. Translating into the notation of the present paper, Hall proposes that $X(\mathrm{u}) = x$ is an actual cause of $Y(\mathrm{u}) = y$ iff the following three conditions are satisfied:

(H1) The actual value of X is x, and the actual value of Y is y, for unit u.

(H2) There exists a manipulation $do(\boldsymbol{W} = \boldsymbol{w})$ for the variables in some subset $\boldsymbol{W}$ of the variables in $\boldsymbol{V} \setminus \{X, Y\}$ where the values in $\boldsymbol{w}$ are all default values of the variables in $\boldsymbol{W}$ and such that $Y = y$ after the manipulation.

(H3) There exists a manipulation $do(X = x^*)$ for $x^* \neq x$ and such that $Y_{X=x^*}(\mathrm{u}) \neq y$ whenever all the variables in $\boldsymbol{W}$ are fixed by the manipulation $do(\boldsymbol{W} = \boldsymbol{w})$ in (H2).

The central idea behind condition (H2) is that the only permissible manipulations are those that set variables to their default values. As Hall writes:

> In one situation, lots of events occur—*that is*, various bits of the world exhibit *deviations from their default states*. In another situation, strictly fewer events occur—*that is*, some of the bits of the world that are in deviant states in the first situation are in their *default states* instead; and every other bit is in the same state as it was. That is what it is for one situation to be, as I will call it, a *reduction* of another. Letting the "null" reduction of a situation just *be* that situation, we can now say the [sic] C causes E iff there is some reduction of the C-E situation in which E depends on C. (129)

In other words, a *reduction* of a structural equation model is the model obtained by setting an arbitrary subset of the variables in the original model to their default values. Conditions (H2) and (H3) require that the value of Y depends on the value of X in some reduction of the original model.

3.2 OVER-DETERMINATION AND ELECTION RESULTS

I begin this section by applying the various theories of actual causation to an example of over-determination. I then consider three voting scenarios: (1) two-candidate, simple-majority

elections *without* abstentions; (2) two-candidate, simple-majority elections *with* abstentions; and (3) three-candidate, simple-plurality elections with or without abstentions. I claim that the deliverances of the theories are prima facie wrong. Hence, these simple voting scenarios provide counter-examples to the theories. I illustrate how the theories go wrong in voting cases with a representative proof. (Complete proofs are relegated to an appendix.)

3.2.1 Over-Determination

Recall the perfect assassins, Ralph and Lauren, from Chapters 1 and 2. Suppose that Ralph and Lauren simultaneously fire rifle shots at King Victim, and both shots hit Victim in the head, killing him instantly. The model has only one equation: $V(u) = R(u) + L(u)$, where '+' is the Boolean OR. The actual values of the variables in this case are $R(u) = 1$, $L(u) = 1$, and $V(u) = 1$. Either shot would have sufficed to kill Victim. Neither Ralph's shot nor Lauren's shot was a difference-maker in the actual circumstances. Manipulating Ralph's shot makes no difference to Victim's state. Neither does manipulating Lauren's shot.

What should one say, then, about whether Ralph's shot or Lauren's shot actually caused Victim to die? Two reactions are possible here: individualist and collectivist.[42] The collectivist denies that the individual over-determining occurrences are actual causes of the over-determined outcome but asserts that the mereological sum of the over-determining occurrences *is* an actual cause of the outcome. On the collectivist view, no special moves are required in order to deal with cases of over-determination, since an over-determined outcome counterfactually depends on the mereological sum of the over-determining occurrences in the actual circumstances.

By contrast, the individualist has the intuition that every over-determining occurrence is individually an actual cause of the over-determined outcome. Schaffer (2003) provides four

[42] See Schaffer (2003) for an excellent review of the problem and an argument that individualism and collectivism exhaust the plausible reactions to over-determination cases.

arguments in favor of the individualist view. The individualist view makes better sense of theoretically important aspects of causation, like prediction, explanation, and attribution of responsibility. The individualist view is naturally consistent with the fact that each over-determining occurrence is connected to the outcome by an independent, completed (causal) process. The individualist view is better able to explain the collective causal powers of the over-determining occurrences. And the individualist view is better able to account for the pragmatics of ordinary causal discourse.

However, the individualist pays a modal cost in order to accommodate the intuition that every over-determining occurrence is an actual cause of the over-determined outcome. Instead of attending to counterfactual dependence only in the actual circumstances, the individualist must attend to counterfactual dependence *in counterfactual circumstances* as well. Consider the assassins again. Victim's death does not counterfactually depend on Ralph's shot or on Lauren's shot in the actual circumstances. But the individualist wants to count both shots as actual causes of Victim's death. In order to see that they are both actual causes, the individualist recommends that we imagine the counterfactual scenario in which Ralph *did not* make an accurate shot. In the *counterfactual scenario* (but not in the actual scenario), Victim's death counterfactually depends on Lauren's shot. (Similar reasoning works for Ralph's shot in the counterfactual scenario in which Lauren's shot is bad.)

All of the theories of actual causation that I have been considering accommodate the individualist intuition by taking into account difference-making in counterfactual scenarios. Every theory of actual causation under consideration allows variables not on a path from the putative cause variable to the effect variable to be set to non-actual values, as long as the new setting does not change the value of the effect and, in Hall's case, as long as the change is to a

default value. In this way, the theories of actual causation recover the individualist intuition. However, the constraints on how non-actual values may be assigned to variables—namely, that the new values may only be assigned to off-path variables and that each new value must be the default value of its respective variable—are too weak or permissive. The constraints do not exclude enough counterfactual scenarios from consideration.

3.2.2 Election Results

Voting scenarios are idealizations of cases important to historians, ethicists, legal theorists, and diagnosticians. Sometimes the idealization is not very noticeable: for example, deciding whether Olympia Snowe's October 2009 vote in the Senate Finance Committee was an actual cause of the health reform bill going to the Senate floor. Sometimes the idealization is extreme: for example, deciding whether the Missouri Compromise was an actual cause of the American Civil War. Voting scenarios—even simple ones—capture a number of interesting cases for ethical and legal theory as well. Are all of the members of successful lynch mobs actual causes of someone being hanged? What about people who actively resist the mob's actions or bystanders who simply watch? Real cases will typically not be as clean and simple as the cases considered below; however, if a theory fails on the simplest of cases, then we should presume against the theory succeeding (in general) for more complicated cases, and when it *does* succeed, we should treat the success as accidental and uninformative.

All of the election scenarios I consider in the present paper share the structure pictured in Figure 3.1 below. In this figure, each vote (or voter) is labeled with a "V_i," and the outcome is labeled "Elect." This greatly simplifies our work, since for any vote, there is only one path to the outcome and that path contains a single directed edge.

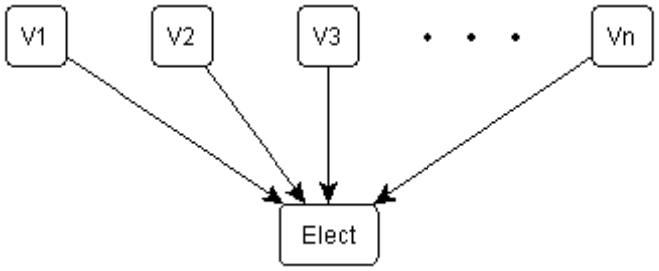

Figure 3.1: Causal Graph for Simple Voting Scenarios

The real work of my examples is done by the number of values the variables are allowed to range over and the non-linearity of the functional relationships in the SEMs representing voting scenarios, rather than by complicated graphical features.[43] In nearly all philosophical discussions of causation, especially where the relata of the causal relation are assumed to be events, SEMs that represent the cases being discussed will involve only indicator variables (binary variables that have values, "Yes, event *e* occurred," or "No, event *e* did not occur"). However, indicator variables are often not well-suited to real work in causal representation and inference, where variables typically have multiple values. The voting cases show that apparently novel problems arise when one considers models having variables with more than two values.

As we have seen, the individualist pays a cost to accommodate the intuition that every over-determining occurrence is an actual cause of the over-determined outcome: the individualist has to attend to counterfactual dependence relations both in the actual circumstances and in counterfactual circumstances as well. The theories run into problems with voting scenarios because they are too liberal in the range of counterfactual circumstances they consider. The deliverances of the various theories with respect to two-candidate, simple-majority elections with

[43] Graphically more complicated voting scenarios can easily be imagined. One might include direct causal dependencies between pairs of votes. One might include common causes of votes (owing to the influence of a demagogue, for example) or common effects intermediate between the votes and the outcome of the election (in order to model voting machines or electors in the Electoral College, for example). Still, I think we ought to get clear about the simplest cases first, for if an account of actual causation cannot get the simplest cases right, it seems unreasonable to hold out hope that it will get more complicated variations right.

abstentions, two-candidate, simple-majority elections without abstentions, and three-candidate, simple-plurality elections are summarized in Table 3.1:

Table 3.1: Voting Scenario Results

	Two-Candidates, No Abstentions		Two-Candidates, with Abstentions			Three-Candidates		
	A	B	A	B	Null	A	B	C
Hitchcock & Woodward	Yes	No	Yes	No	Yes	Yes	Yes	Yes
Halpern & Pearl	Yes	No	Yes	No	Yes	Yes	Yes*	Yes
Hall	N/A	N/A	Yes	No	Yes	Yes	Yes†	Yes†

* Yes, except when the number of votes for candidate *A* is strictly less than twice the number of votes for candidate *B* plus one, the total number of votes is odd, and there are no votes for candidate *C*.
† Yes, as long as there is at least one vote for each of the candidates.

The table describes the three types of election being considered. Under each type of election is a listing of the choices available in that election: vote for A, vote for B, vote for C, or abstain (Null). Without loss of generality, the candidates are assumed to be listed in descending order of the number of votes they received. Specifically, candidate A is assumed to have received the most votes, and candidate B is assumed to have received at least as many votes as candidate C. Each row of the table marks a specific theory of actual causation. If a row contains a "Yes" entry in the column for candidate X (or Null), then the theory counts every vote for candidate X (or every abstention) as an actual cause of the result of the election, with the provisions indicated in the notes to the table.

Two quick counter-examples may be constructed from these results. I leave extended discussion and further examples to Section 3.2. *First Counter-Example.* Suppose Jack and Jill live in a Congressional district that is overwhelmingly Republican. Jack and Jill both know that the Republican candidate is going to win the election (and, in fact, the Republican does win). Jill prefers the Democratic candidate, but she finds herself so disgusted with the first-through-the-

gate election system that she abstains in protest. Jack, on the other hand, prefers the Republican candidate. Since Jack believes that the Republican will win easily without his vote and since he has lots of important things to do, Jack decides to abstain. The theories treat Jack's abstention and Jill's abstention as exactly alike, but they are clearly different. *Second Counter-Example*. Elizabeth's teacher is letting her students vote on whether to read *Pride and Prejudice* or *Sense and Sensibility* for their section on Jane Austen. The class votes 18-2 in favor of *Pride and Prejudice*, with Elizabeth on the losing side. Elizabeth does not like *Pride and Prejudice*, and she likes losing even less. So, she asks her teacher if the class can vote again with *Emma* included in the options. The teacher agrees, and the class votes 18-1-1 in favor of *Pride and Prejudice*. The theories say that not only has Elizabeth lost in the second vote, she actually caused *Pride and Prejudice* to be selected as the book for the Jane Austen section of her class.

The results for three-candidate, simple-plurality elections are easily extended to elections having four or more candidates with or without abstentions. All the theories endorse the claim that every vote and every abstention (with the exceptions already noted) is an actual cause of the result of a simple-plurality election with three or more candidates. The proofs of the results summarized in Table 1 exploit over-permissiveness about how the votes may be redistributed in counterfactual situations. In order to save space and not distract the reader with irrelevant details, I will only exhibit one simple, illustrative proof in the main text. I have left the other proofs to the appendix or as exercises for the reader.

3.2.3 An Illustrative Proof

I prove that for Hitchcock and Woodward, every vote in a three-candidate, simple-plurality election is an actual cause of the result of the election, whatever that result might be. Consider an election in which there are $2k$ votes. (The result is identical for elections involving an odd

number of votes. The proof for the odd-numbered condition is left as an exercise for the reader.) Suppose that i votes are for candidate A, j votes are for candidate B, and l votes are for candidate C. Further, suppose without loss of generality that $i > j \geq l$.

To see that every vote for candidate A is an actual cause of candidate A's victory, choose a vote $V_A = A$. Distribute the votes such that there are $k + 1$ votes for candidate A (including V_A), $k - 1$ votes for candidate B, and no votes for candidate C. Changing the value of V_A from A to B results in a tie (k votes for A against k votes for B). Hence, $V_A = A$ is an actual cause of Elect = A.

To see that every vote for candidate B is an actual cause of candidate A's victory, choose a vote $V_B = B$. Distribute the votes such that there are k votes for A, one vote for B, and $k - 1$ votes for C. Changing the value of V_B from B to C results in a tie (k votes for A against k votes for C), so $V_B = B$ is an actual cause of Elect = A. Similarly, every vote $V_C = C$ is an actual cause of Elect = A.

Thus, according to Hitchcock's theory and according to Woodward's theory, every vote cast in an election having three (or more) candidates is an actual cause of the result of the election! The proof works because Hitchcock and Woodward's theories always allow votes to be re-distributed (in the counterfactual condition) such that the election is decided by a single vote and such that the single vote is for the candidate we are thinking about. Often, the theory allows a full third or more of the votes to be assigned arbitrary new values in the counterfactual condition. The result in no way depends on the actual number of votes cast for each candidate. Even if no one votes for candidate C, every vote for candidate B is an actual cause of candidate A's victory.

3.3 INTUITIONS

I now want to explore the deliverances of the theories of actual causation with respect to the simple voting scenarios I have been considering and suggest some generic revisions. I point out a number of situations where the deliverances of theory do not accord well with my intuitions. I identify several intuitions and cast them as constraints on a theory of actual causation. I then argue for two things: (1) what counts as an actual cause depends on *conditional* defaults, and (2) what makes something an actual cause is not an intrinsic feature of that thing.

3.3.1 Symmetry and Causal Production

In the case of two-candidate, simple-majority elections without abstentions, the theories say that all and only the votes for the winning candidate are actual causes of the winning candidate's victory. Insofar as one has individualist intuitions, intuition appears to agree with the theories in this case. One reason to be an individualist about elections is that each vote is a *producer*. If an election involved only one vote, then that vote would produce or determine the result. Votes for candidate *A* produce victory for candidate *A*, assuming that candidate *A* wins the election. Votes for candidate *B* produce victory for candidate *B*. Since the winning votes are indistinguishable (i.e. all of the votes have equal weight in determining the outcome), symmetry requires that all of the votes be considered equally efficacious.

3.3.2 Abstentions, Contrasts, and Conditional Defaults

In the case of two-candidate, simple-majority elections where voters are allowed to abstain, the theories say that all votes for the winning candidate and all abstentions are actual causes of the winning candidate's victory (and no other votes are actual causes of the winning candidate's victory). Counting the votes for the winning candidate as actual causes makes sense in the same

way that it made sense for the case without abstentions. But how should we understand the causal role of abstentions?

One thing is clear about the causal role of abstentions in voting scenarios: if abstentions have any causal role at all, they have it in virtue of facts about causal dependence, not in virtue of facts about causal production.[44] Abstentions (and absences generally) do not *produce* anything. From the perspective of causal production, we should be able to ignore abstentions altogether: an election in which candidate *A* receives six votes, candidate *B* receives three votes, and two people abstain is equivalent to an election in which candidate *A* receives six votes, candidate *B* receives three votes, and seventeen people abstain.

However, those two elections strike me as being importantly different. In the second case, but not in the first case, the abstentions seem to matter. My initial impression is that in order for any abstention to matter, it must be the case that the abstentions matter collectively. In other words, if an abstention is to count as an actual cause of the result of an election between two candidates, the number of abstentions must be greater than or equal to the gap between the winner and the loser. This intuition turns Schaffer's discussion of the source of collective causal powers on its head.[45] I ask, "How could the *individual* abstentions matter if the collection of abstentions does not?" When the actual votes are six for *A*, three for *B*, and two abstaining, I want to say that the abstentions, as a group, do not matter. Therefore, I want to say that the individual abstentions do not matter. However, when the actual votes are six for *A*, three for *B*, and seventeen abstaining, I want to say that the abstentions as a group do matter, or at least, they might matter. So following Schaffer, I want to say that the individual abstentions (might) matter.

[44] See Hall (2004) on the differences between causal production and causal dependence.
[45] See Schaffer (2003), Section 5.

The individual abstentions *might* matter, but my confidence about the judgment that they *do* matter depends on what the voters who abstained would have done—who they would have voted for—had they decided (or been forced) to vote. For example, if everyone who abstained would have voted for candidate *A* (the winning candidate), then I would say that the abstentions were not actual causes of the result of the election. Rather, candidate *A* won despite the fact that some *A*-supporters abstained. On the other hand, if everyone who abstained would have voted for candidate *B* (the losing candidate), then I would say that the abstentions were actual causes of the result of the election (assuming that the number of votes for *B* plus the number of abstentions is greater than or equal to the number of votes for *A*). Abstentions count as actual causes or fail to count as actual causes in the light of some relevant contrast. Two examples will make this clearer.

Imagine that in a certain department, a search committee recommends job candidates one at a time and then the faculty votes on whether to hire the job-seeker. Professors may vote "Yes," "No," or "Abstain." If the vote is tied, then the question is tabled for two days of debate and then brought to a new vote. Now, imagine that the committee has recommended a controversial candidate named Steve. Dr. Smith has an unfavorable impression of Steve, but at the last minute, he decides to abstain instead of voting "No." The vote comes in at 4-3 in favor with Dr. Smith abstaining, and Steve is offered a job. I have the clear intuition that Dr. Smith abstaining *rather than voting* "*No*," caused Steve to be offered the job. But now imagine that two other professors, Dr. Crane and Dr. King, who had favorable impressions of Steve also decided to abstain. With the final vote at 2-3 against and three abstaining, Steve is not offered a job. I have the clear intuition that by abstaining, Dr. Crane and Dr. King prevented Steve from being offered the job. But I do *not* want to say that Dr. Smith's abstaining prevented Steve from

being offered the job; rather, I want to say that Steve was offered the job *despite* Dr. Smith's abstention.[46]

In the 2008 U.S. Presidential election, many out-of-state university students in Virginia and Colorado were told that they could not vote in their school's state.[47] This was false. Since students overwhelmingly support the Democratic Party, had John McCain won the election, it might fairly have been said that many students not voting was an actual cause of McCain's victory. On the other hand, since Obama won the election, student abstentions were not actual causes of anything.[48] Judgments about actual causation are contrastive—something hidden by the way actual causation is treated by the theories under consideration.[49] When I say that Jill's abstention was an actual cause of the Republican candidate winning the election, I am saying that Jill's abstaining rather than voting for the Democratic candidate was an actual cause of the Republican candidate winning the election.[50] Where do the relevant contrasts come from? I claim that they come from facts about *conditional defaults*.

But first, we need to rethink the notion of default. Hall's conception of defaults comes out of the dominant view that causation is a relation between events, where an event is binary—

---

[46] Whether you think that the various professors are responsible for Steve getting or not getting the job is a different, though related question. See Livengood and Machery (2007) for a discussion of causation by absence and its relation to explanation. See Sartorio (2004) for a discussion of the relationship between causation and ascriptions of moral responsibility.

[47] See the New York Times article here http://www.nytimes.com/2008/09/08/education/08students.html and the Colorado Springs Gazette article here http://www.gazette.com/articles/vote-40925-colorado-students.html.

[48] Perhaps you think, like my friend and colleague Justin Sytsma, that if McCain had won the election under these circumstances, then *the voter fraud* perpetrated in Colorado and Virginia would have been the actual cause of his victory. I agree that it would have been an actual cause in such circumstances. Moreover, it would have been the most salient actual cause from the perspective of moral and legal responsibility attribution. And, if a variable for fraud is included in the structural model, the theories will all say that the fraud was an actual cause of the outcome. However, the theories will continue to say the same things about the votes that they said before; adding a variable for voter fraud will not change the deliverances of the theories with respect to the causal role of the votes. Actual causation is widely agreed to be non-transitive. If $A(u) = a$ is an actual cause of $B(u) = b$ and $B(u) = b$ is an actual cause of $C(u) = c$, it might still be the case that $A(u) = a$ is not an actual cause of $C(u) = c$. On the other hand—and here is where the problem comes in—if the structural model looks like $A \rightarrow B \rightarrow C$, then every theory of actual causation supposes that if $A(u) = a$ is an actual cause of $C(u) = c$, then it must be the case that $A(u) = a$ is an actual cause of $B(u) = b$ and $B(u) = b$ is an actual cause of $C(u) = c$.

[49] Hitchcock (personal communication) reminded me of the importance of contrasts in causal judgments.

[50] I am making the simplifying assumption that there are only Republicans and Democrats in the election.

something that either happens or does not happen. From this perspective, it makes sense to say that by default, nothing happens. The default is for whatever event actually occurred to have not occurred. Hall thinks about this operation as removing an event and leaving a void in its place. By contrast, I claim that the default for any occurrence is exactly what actually happened. If Jill actually abstains in an election, then the default is for Jill to have abstained. If Jill actually votes for a Democratic candidate, then the default is for Jill to have voted for a Democratic candidate. Given the defaults, we can ask a further question: how would Jill have voted if she had been prevented from doing what she did by default? In other words, we can ask about Jill's conditional defaults. Suppose Jill actually abstained in some election. The question, "How would Jill have voted had she not been allowed to abstain?" is a question about a conditional default for Jill. Such questions may be iterated, provided there are enough options available. If an election involves six candidates and allows abstentions, then we might ask how Jill would have voted had she been prevented from abstaining and also prevented from voting for either candidate *A* or candidate *B*.

Define the zeroth-order default for a variable *V* as the value that *V* takes on in the actual circumstances, and define the $(n + 1)$th-order default for a variable *V* as the value that *V* would take on were it prevented from taking on any lower-order default value. Conceptualizing defaults and actual causation in this way requires some rethinking of the $do(\cdot)$ calculus. Ordinarily, the $do(\cdot)$ operator forces its target to take on a specific value (or when probabilities are involved, a specific distribution). Thus, the $do(\cdot)$ operator may be understood as completely constraining its target variable. On the present view of defaults, a more generic partial-constraint operator is required. Instead of writing $do(X = x)$, write $do(X \in \{x_i\})$ for admissible values x_i.

The relevant contrasts for a theory of actual causation come from considering what an agent (or other target of investigation) would do by default (by its nature), *conditional on* some constraint. When one asks about whether some occurrence was an actual cause of some outcome without specifying a conditioning constraint, then one is asking whether that occurrence rather than its first-order default was an actual cause of the outcome. Determining the conditional defaults will often require consultation of some special science for its resolution. Philosophers of science (and especially philosophers of physics) often say that metaphysics must take stock of physics. In voting cases, it appears that metaphysics must take stock of psychology, neuroscience, and much else as well.

3.3.3 Asymmetry, Stability, and Irrelevant Details

In the case of elections involving three or more candidates (with or without abstentions), every theory of actual causation under consideration counts *every* vote (and *every* abstention) as an actual cause of the winning candidate's victory, with the rare exceptions mentioned in footnotes to Table 3.1.

On its face, this result is absurd. How could a vote *against* a candidate possibly count as an actual cause of that candidate being elected? Two relatively recent U.S. Presidential elections show how votes for third-party candidates might count as actual causes of the outcome of three-candidate simple-plurality elections. For instance, when Clinton defeated Bush in 1992, some pundits suggested that Perot had siphoned off enough votes from Bush to give the election to Clinton. Similarly, when Bush defeated Gore in 2000, some suggested that Nader cost Gore the victory. However, I would be very surprised to hear anyone say that *Gore cost Nader* the 2000 election or that *Bush cost Perot* the 1992 election. My attitudes toward the two losing candidates in three-candidate, simple-plurality elections are not symmetric. Only the last-place finisher

seems fit to count as a cause of the victorious candidate winning the election. The intuition extends to the voters for the candidates. I have witnessed many conversations recently in which one person expresses an interest in voting for a third-party candidate, and his or her interlocutor replies, "If you vote for a third party and the Democrats lose, it will be your fault." (Given the political preferences of the people in these conversations, the person voting for a third party would not be considered blameworthy if the Democratic candidate won.) Call this the *asymmetry intuition*.[51]

None of the theories of actual causation considered in this paper respects the asymmetry intuition. Hitchcock's theory and Woodward's theory also fail to capture two further intuitions, which I call the *strong and weak stability intuitions*. According to the weak stability intuition, what counts as an actual cause of the outcome of an election in which some candidate receives no votes should be the same as what counts as an actual cause of the outcome of an otherwise identical election in which the candidate that received no votes does not appear. Here is an example. Suppose a corporate board consisting of 23 members takes a vote to decide whether to build their new facility in New York or Los Angeles. The vote is 16 for New York and 7 for Los Angeles. In this case, Hitchcock's theory and Woodward's theory both tell us that the 16 votes to build in New York are actual causes of the company building their new facility in New York, while the other 7 votes are not. Now, suppose that one of the board members thinks the board should consider building in Chicago, though she herself does not think Chicago is really the best place to build. She presents this view to the board, and consequently, the members vote on

[51] Hitchcock (personal communication) replies to the asymmetry intuition as follows: "I agree that it is counterintuitive to say that a vote for Gore counts as a cause of Bush's victory. But I'm not sure that is the same as saying that Gore cost Nader the election. We can imagine scenarios where that might sound true, even if the votes are the same. E.g., at first Nader and Bush are the only candidates, and slightly more voters favor Nader. Then Gore enters. Most Nader voters switch to Gore, but no Bush voters do. Bush narrowly wins. In this case, I don't think it's so unreasonable to say that Gore cost Nader the election. If Gore hadn't run, Nader would have won." I am sympathetic to this reply; however, I think it adds something to the structure of the election scenarios by bringing in time.

whether to build in New York, Los Angeles, or Chicago. The vote is 16 for New York, 7 for Los Angeles, and none for Chicago. According to the weak stability intuition, what counts as an actual cause of the company building in New York in this scenario should be the same as what counts as an actual cause of the company building in New York in the previous scenario. However, according to Hitchcock and Woodward's theories, in the second scenario, the seven votes to build in Los Angeles are actual causes of the company building its new facility in New York!

According to the strong stability intuition, what counts as an actual cause of the result of an election should be insensitive to partitioning the votes for a losing candidate among the losing candidate and some number of new candidates (where by "new" I mean candidates not on the ballot in the original election). In the previous example of the corporate board, the strong stability intuition says that if the seven votes for Los Angeles were re-distributed as votes for Los Angeles and Chicago (but the 16 votes for New York were left alone), then the actual causes should be the same as in the original case. Similarly, if the board decided to add Chicago and Houston for consideration and the new vote came out as 16 for New York, 4 for Los Angeles, 2 for Chicago, and 1 for Houston, then according to the strong stability intuition, only the votes for New York should count as actual causes of the company building its new facility in New York.

Hall's theory fails to capture the strong stability intuition but not the weak stability intuition. Halpern and Pearl's theory fails to capture both the weak and strong stability intuitions, though not for all cases. One might think that this is good news for Halpern and Pearl, but it is not. Halpern and Pearl's theory will not count votes to build in Los Angeles as actual causes of the company building in New York, when the 23-member corporate board votes 16-7-0 in favor of building in New York over Los Angeles and Chicago. However, the reason is very

peculiar. In this case, the total number of votes is odd, there are no votes for the third-place option, and the winning option has more than twice as many votes as the second-place option. Under those conditions, Halpern and Pearl's theory only counts the first-place votes as actual causes of the outcome of the election. However, if any of those conditions is different, then Halpern and Pearl's theory will count all of the votes as actual causes of the outcome of the election. For example, if the corporate board has 24 members that initially voted 17 to 7 for New York, then it endorses the same conclusion that Hitchcock and Woodward's theories endorse. However, the conclusion that votes for Los Angeles are not actual causes when Chicago is not a live option but are actual causes when Chicago is a live option—even if no one picks that option—is no more compelling when the board has an even number of members than it was when the board had an odd number of members. In virtue of paying attention to whether the total number of votes is even or odd, Halpern and Pearl's theory fails to capture another important intuition, which I call the *irrelevant details intuition*.

Any process that keeps track of overall counts of occurrences will work out similarly to voting. For example, imagine that the area around an apple tree is divided into two patches (*A* and *B*), and suppose that we count how many apples fall from the tree and come to rest on each patch. Suppose that more apples land on patch *A* than on patch *B*. All of the theories under consideration will say that all and only the apples that landed on patch *A* were actual causes of patch *A* being covered with more apples. However, if we subdivide patch *B* into two new patches *C* and *D*, then (with the exceptions already noted), every theory will say that all of the apples were actual causes of patch *A* being covered with the most. The theories pay no attention to how likely each apple was to land on a given patch. Nor do they pay attention to how the apples were actually distributed over the available patches.

Again, the problem is that these theories of actual causation are too permissive in the range of counterfactual scenarios they consider. Instead of allowing arbitrary re-distributions of the values of variables in a structural model, theories of actual causation ought to consider only re-distributions in line with first-order defaults, unless a specific contrastive question is at stake. Hence, in order for a third-party vote to count as an actual cause of the winning candidate's victory, there should be a counterfactual scenario—constructed by re-distributing votes according to their first-order defaults—in which the third-party vote is a difference-maker.

3.4 HOW BAD IS IT?

As with many philosophical theories, the theories of actual causation considered in this chapter are based on simple, intuitive ideas. They then add bells upon whistles in order to account for places where our intuitions diverge from the theory. My voting examples led me to add another whirligig. But how bad is the failure that motivates the new noise-maker, really?

The answer to that question depends on how we agree to adjudicate between competing theories of actual causation, which itself depends on what theories of actual causation are supposed to be doing for us. On the one hand, a theory of actual causation could be purely descriptive—telling us how we *do* reason about actual causation. In that case, the proper tool is behavioral experimentation. On the other hand, a theory of actual causation could be normative—telling us how we *ought* to reason about actual causation. In that case, it is not at all clear that ordinary reasoning about actual causation is a good guide to theorizing. Just as we do not think that widespread commission of the gambler's fallacy makes that form of reasoning normatively correct, ordinary people might reason counter-normatively about actual causation.

Many philosophers have accepted the demand that theories of (actual) causation satisfy ordinary causal intuitions or attributions. However, it is not clear that theorists should accept that demand. On the one hand, many philosophers have been skeptical about the use of intuitions in metaphysics. Even if one does not harbor any in-principle objections to intuition-mongering, Glymour et al. (2010) present a compelling argument that the method of cases is practically hopeless for theorizing about actual causation. However, an alternative method for adjudicating between competing theories of actual causation—in the normative sense—has not been provided. In the remainder of this dissertation, I consider the prospects for experimental approaches to theorizing about actual causation.

CHAPTER FOUR

EXPERIMENTS

How do ordinary people reason about actual causation and what are they trying to accomplish by reasoning about actual causation? In this chapter, I consider some recent experimental work on actual causation. My starting place is Hitchcock's (2007a) theory connecting actual causation to the principle of sufficient reason.[52] The theory fails for want of an adequate account of norms, which is made clear by experiments I conducted with Justin Sytsma and David Rose. Recently, Hitchcock and Knobe have suggested an alternative approach to the relationship between causation and norms, which might be read as a patch of Hitchcock's 2007 theory. The new theory is also inadequate, but for different reasons.

Christopher Hitchcock (2007a) proposes an elegant theory that specifies circumstances in which the counterfactual dependence of an event *e* on another event *c* is both necessary and sufficient for the event *c* to count as an actual (or token) cause of the event *e*. The theory has two main components and a commitment. The two components are (1) an account of how to get actual causes from a graphical model and a specification of default values for the variables in the model and (2) a background theory that specifies the default values for the variables in the model. The commitment is to faithfully represent ordinary causal attributions, i.e. to correctly capture what ordinary people say about what causes what in various cases. I show by

[52] Hitchcock's 2007 theory is sort of a revision of his 2001 theory and sort of not. The 2007 theory does not apply to all cases. Where it does apply, Hitchcock thinks that it is consistent with his 2001 theory, and where it does not apply, Hitchcock tells us to rely on the 2001 theory. However, the 2007 theory deploys resources—connected to a notion of "default"—that are not used by the 2001 theory.

experiment that the two components—as Hitchcock has stated them—and the commitment to correctly capture what ordinary people say about causation are not mutually compatible. Something has to go, and I argue that what ought to go depends on the goal. If the goal is to predict what people will say about causation, then the background theory needs to be radically altered. If the goal is to do good diagnostic work, then the commitment to correctly capture what ordinary people say about causation ought to be dropped. I argue that Hitchcock (2007b) is wrong to say that the metaphysical concept of causation is a dispensable hybrid of the scientific and folk-attributive concepts.

Whenever an individual asserts that, denies that, or expresses indifference about a claim that something or someone caused (or was a cause of) some outcome, that the individual is making a *causal attribution*. According to Hitchcock, a theory of actual causation should accurately describe ordinary, untutored causal attributions for a broad range of cases. Call these untutored attributions *folk causal attributions*, and call the commitment to accurately describe folk causal attributions the *folk attribution desideratum* (*FAD*). Many philosophers have expressed a commitment to the FAD. For example, Lewis (1986, 194) writes: "When common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree. If an analysis of causation does not deliver the common-sense answer, that is bad trouble."

One might object that philosophers are not really interested in folk causal attributions but in folk causal judgments or intuitions. Perhaps the answers delivered by common sense in Lewis' statement above are only answers in the head, not answers actually delivered by people. I am confident that Hitchcock, at least, is committed to the FAD, and I argue for this claim in Section 4.1 below. But, setting that aside for the moment, I claim that since causal judgments

underlie causal attributions, the causal attributions one makes provides evidence about the causal judgments one makes or the causal intuitions one has.[53] I grant that causal judgments are often difficult to read off from causal attributions. For example, what people *say* with respect to what causes what might very well be biased by conversational pragmatics. However, if causal judgments are to make any difference or have any place in philosophical theories, they have to have some influence on causal attributions. So, even if one has a weaker commitment, say to a *folk intuition desideratum* (*FID*), some account must be given of folk causal attributions.[54]

In the present chapter, I provide experimental evidence that Hitchcock's (2007a) theory of actual causation does not satisfy the FAD. I argue that whether or not the FAD should be retained depends on the goal one has. I suggest two different goals and argue that only one of those goals makes commitment to the FAD desirable. Moreover, I argue that Hitchcock (2007b) was wrong to think that the metaphysical concept of causation is dispensable.

Here is how I proceed. In Section 4.1, I show that Hitchcock is committed to the FAD. In Section 4.2, I present Hitchcock's theory of actual causation. In Sections 4.3 and 4.4, I provide experimental evidence that Hitchcock's theory does not satisfy the FAD. Finally, in Section 4.5, I consider what parts of Hitchcock's theory must be abandoned and why.

4.1 HITCHCOCK AND THE FOLK ATTRIBUTION DESIDERATUM

Hitchcock (2007b) discusses three concepts of causation: scientific, folk attributive, and metaphysical. Internal coherence requires that Hitchcock take his theory of actual causation to be a theory of the folk attributive concept of causation. Hence, the FAD is a natural commitment

[53] I take this claim to be something of a triviality. As Hacker (2009) remarks about several prominent twentieth-century Anglophone philosophers, "Most would have agreed that a primary way to attain a clear view of our concepts was to investigate the use of words that express them" (338).
[54] Thus, explaining-away projects like those of Weatherson (2003) and Ichikawa (2009) really need to have two layers, one at the level of intuitions themselves and one at the level of attributions.

to have. I provide evidence internal to Hitchcock (2007a) that indicates he actually is committed to the FAD.

The argument is simple. First, Hitchcock's theory of actual causation cannot be a theory of the scientific concept, because unlike causal relations under the scientific concept, actual causal relations cannot be simply read off from a causal model. Second, Hitchcock's theory of actual causation cannot be a theory of the metaphysical concept, since Hitchcock (2007b) claims that the metaphysical concept of causation is an unstable and dispensable hybrid of the scientific and folk-attributive concepts, and it would be pragmatically inconsistent to offer a metaphysical account of causation while simultaneously claiming that no one should be offering metaphysical accounts of causation. Finally, Hitchcock's theory of actual causation is congruent with the folk-attributive concept. Both are retrospective. Both are (or could be) employed to facilitate expressions of praise and blame. And both are discriminatory.[55] Given that Hitchcock is attempting to characterize the folk attributive concept of causation, requiring that his theory satisfy the FAD is a plausible demand, and it is a demand that Hitchcock accepts.

Let us take the components of our argument one at a time. Regarding the scientific concept, Hitchcock writes, "The scientific concept of causation is the focus of the recent literature on causal modeling. Physical systems possess a causal structure that can be represented by a *causal model*. … This conception of causation is scientific in the sense that the causal model generates a set of predictions about the system in question" (2007b, 510). In fact, the scientific concept of causation appears to be entirely exhausted by the parameterized causal structures described by causal models. The scientific concept of causation is exhausted by facts about causal structure.

[55] A theory is discriminatory if it distinguishes causes from mere background conditions; it is non-discriminatory if it does not. The terminology is due to Lewis (1986, 162). See Hitchcock and Knobe (forthcoming) for an extended discussion.

However, Hitchcock notes that while a causal model may be used to represent the causal structure of some specific episode, "when a causal model is used to characterize the causal structure of some specific episode, it will not in general be possible to 'read off' the relations of token causation from the causal model" (511). As Hitchcock (2007a) and Hitchcock (2009) make clear, an adequate theory of actual causation must supplement the bare equations in a causal model somehow—Hitchcock's own approach being to privilege some values of the variables in the model as defaults.[56] Hence, Hitchcock's theory of actual causation goes beyond the scientific conception of causation.

In order to apply Hitchcock's theory, one has to know some facts about the relevant causal structure *and also* some facts that pick out the default states of the variables in that structure. Importantly, causal structure by itself does not determine the default values of the variables in the structure. Moreover, Hitchcock's theory of actual causation does not generate a set of predictions *about the causal system being modeled*. Instead, Hitchcock's theory makes predictions about what people will say or think about a given causal system, which is described by a causal model. Hence, Hitchcock's theory of actual causation is not scientific in the same sense that the scientific concept of causation is.

By contrast to the scientific concept of causation, the folk attributive concept of causation makes no predictions about future regularities or the results of interventions on a given causal system. Rather, the folk attributive concept is the concept deployed by ordinary people in making everyday judgments about causal responsibility. Hitchcock (2007b) writes:

> I take the term 'attributive' from Hart and Honoré, who distinguish the types of causal claim made in 'scientific' inquiries from those made in 'attributive' inquiries. It is the latter type of claim, they assert, that is of primary concern to 'the lawyer, the historian, and the plain man'. I have added the modifier 'folk' in order to emphasize that it is

[56] Hall (2007) also argues that causal structure does not determine relations of actual causation, and he also appeals to a notion of default values of a variable in setting out an alternative theory.

> primarily the 'plain man's' conception of causation that I have in mind here. It is just this concept that the 'plain man' applies when he asserts, after the occurrence of some event, that some preceding event was the cause, or a cause, of it. Such causal attributions play a central role in assigning moral or legal responsibility for the event in question. (511-512)

The folk attributive concept of causation accords well with Hitchcock's theory of actual causation. Both are retrospective—they assume an effect event has occurred and then look for the cause of that event. When causal questions are at stake for lawyers and historians, they are questions about what caused some given event or events in the past. And Hitchcock's examples of actual causation follow suit. For example, he writes, "By 'token causation' I mean the sort of causal relationship that is reported in claims such as 'Assassin's poisoning the coffee caused Victim to die'" (2007a, 496). Furthermore, the technical conditions of Hitchcock's theory allow one to judge whether a *specified* effect event counterfactually depends on some event(s) that preceded it; hence, the theory has an essential element of retrospection in it.[57]

Both the folk attributive concept of causation and Hitchcock's theory of actual causation are (or could be) employed to facilitate expressions of praise and blame. We have already seen the connection between the folk attributive concept of causation and judgments of responsibility. In the same vein, Hitchcock (2007a) writes, "Token causation is involved specifically in our post hoc evaluations of responsibility: after the fact, which agent's actions were responsible for the outcome? This notion is important to philosophers since it plays a role in concepts like moral responsibility and singular explanation" (504).

Finally, both are discriminatory. When the "plain man" (or even the far-less-plain lawyer or historian) makes a causal judgment in the attributive mode, he usually picks from a field of active causal processes some single salient event or action and calls that the cause. Hitchcock

[57] One might object that Hitchcock's theory would allow one to say hypothetically whether some events will be judged to have caused some effect event in advance of any of these events occurring. Not only has this maneuver not been explicitly contemplated in the actual causation literature, it would not discharge the essentially retrospective nature of actual causation judgments.

(2007b) writes, "We count the careless tossing of the cigarette as a cause of the forest fire, whereas the oxygen in the atmosphere, while equally necessary for the occurrence of the fire, is relegated to the status of a mere background condition" (512). Hitchcock's theory of actual causation also distinguishes causes and background conditions. He even uses the same example of a cigarette causing a forest fire (2007a, 514).

So far, I have shown that Hitchcock's theory of actual causation is not an account of the scientific concept of causation, and it closely resembles the folk attributive concept of causation. However, Hitchcock (2007b) points out that the metaphysical concept is similar to the folk attributive concept in a number of ways, including its logical form. Is Hitchcock's theory of actual causation meant to be a contribution to metaphysics, perhaps? I think not. Hitchcock explicitly says that we can get by *without* the metaphysical concept of causation. Summarizing his view, he writes:

> We make use of causal knowledge to make predictions, and to guide our interventions in the world. For this purpose, it suffices to have knowledge of the causal structure of the scientific conception. We also make use of causal knowledge when we assign praise and blame for outcomes. For this purpose, the folk attributive concept of causation, value-laden as it is, is entirely appropriate. Thus, in characteristic applications of causal notions, we can make do without the metaphysical concept altogether. (2007b, 515)

Moreover, he thinks that we *ought* to do so. Hitchcock argues that the metaphysical concept resulted from a mistaken belief that whatever objective property in the world grounds folk attributive judgments of causation has the same logical form as the judgments themselves. Thus, he writes, "Metaphysical causation, it turns out, is an unstable compromise between the scientific and folk attributive concepts of causation: it seeks to retain the logical structure of the folk attributive concept while retaining the objectivity of the scientific concept" (2007b, 514).

It would be very odd for Hitchcock to so utterly reject the metaphysical conception of causation and at the same time want to offer a metaphysical theory of causation. For the sake of

pragmatic consistency, Hitchcock's theory of actual causation must be about either the scientific or the folk attributive concept of causation. But I have already argued that it is *not* about the scientific concept of causation. Therefore, it must be about the folk attributive concept. It would be straightforwardly absurd to think that an account of the folk attributive concept of causation need take no notice of the causal attributions ordinary people actually make. Hence, given his other commitments, Hitchcock ought (naturally) to be committed to the FAD. Some textual evidence supports this contention. Hitchcock (2007a, 504-507) notes that his theory of actual causation is not entirely objective owing to its reliance on the distinction between default and deviant values of a variable—more on this distinction below. He considers the possibility of making his theory of actual causation entirely objective by appealing to fundamental laws and using those laws to determine the default and deviant values of the variables in the causal model. However, he rejects this move, writing, "Perhaps a case could be made for allowing only genuine laws of nature to determine default values of variables, *but if we disallow folk theories, we are not likely to arrive at a theory that accords with folk intuitions*" (2007a, 506 emphasis added). The consequent of the conditional in this quotation is supposed to provide reason to reject the suggestion in the antecedent. Hitchcock wants to allow folk theories to determine default and deviant values precisely because he wants to arrive at a theory that respects what people say about actual causation.

4.2 HITCHCOCK'S THEORY OF ACTUAL CAUSATION

Hitchcock presents a condition (TC) specifying the circumstances in which counterfactual dependence of one event *e* on another event *c* is necessary and sufficient for *c* to count as an actual cause of *e*. TC is given in terms of what Hitchcock calls *self-contained networks*, the definition of which requires a distinction between the default and deviant values of a variable.

4.2.1 Default and Deviant Values

Hitchcock thinks that the difference between the default state of a system and deviations from that default is an important component of a correct theory of actual causation.[58] The difference, he claims, is fairly straightforward in most cases, but it is difficult to state precisely. Roughly, the default state of a system is its natural state—the state the system is usually in unless something has been done to it. How the defaults are picked out is a matter for the relevant scientific and folk theories. Hitchcock (2007a) writes:

> As the name suggests, the default value of a variable is the one that we would expect in the absence of any information about intervening causes. More specifically, there are certain states of a system that are self-sustaining, that will persist in the absence of any causes other than the presence of the state itself: the default assumption is that a system, once it is in such a state, will persist in such a state. (506)

He goes on to provide rules of thumb for distinguishing between default and deviant values of a variable. He writes:

> Temporary actions or events tend to be regarded as deviant outcomes. In the case of human actions, we tend to think of those states requiring voluntary bodily motion as deviants and those compatible with lack of motion as defaults. In addition, we typically feel that deviant outcomes are in need of explanation, whereas default outcomes are not necessarily in need of explanation. Frequently, but not always, my deviant values correspond to positive events, and defaults correspond to absences or omissions. (507)

This distinction allows Hitchcock to reformulate Leibniz's Principle of Sufficient Reason, which he then applies to SEMs to derive the following rule: no variable takes on a deviant value if all of its parents take on default values.

4.2.2 Self-Contained Networks and Token Causation

In order to present his rule precisely, Hitchcock defines two new technical notions: a causal network and a self-contained causal network. A causal network is a set of variables in a structural equation model picked out by a graphical condition:

[58] He is not alone. See Hall (2007) for another take on the notion of defaults.

> CN: Let <$\boldsymbol{V}$, $\boldsymbol{E}$> be a causal model, and let X, $Y \in \boldsymbol{V}$. The *causal network* connecting X to Y in <$\boldsymbol{V}$, $\boldsymbol{E}$> is the set $\boldsymbol{N} \subseteq \boldsymbol{V}$ that contains exactly X, Y and all variables Z in $\boldsymbol{V}$ lying on a path from X to Y in <$\boldsymbol{V}$, $\boldsymbol{E}$>. (509)

The notion of a self-contained causal network augments the graphical condition by appeal to the default values of the variables in the network:

> SCN: Let <$\boldsymbol{V}$, $\boldsymbol{E}$> be a causal model, and let X, $Y \in \boldsymbol{V}$. Let $\boldsymbol{N} \subseteq \boldsymbol{V}$ be the causal network connecting X to Y in <$\boldsymbol{V}$, $\boldsymbol{E}$>. Then the causal network $\boldsymbol{N}$ is *self-contained* if and only if for all $Z \in \boldsymbol{N}$, if Z has parents in $\boldsymbol{N}$, then Z takes a default value when all of its parents in $\boldsymbol{N}$ do (and its parents in $\boldsymbol{V} \setminus \boldsymbol{N}$ take their actual values). (510)

Suppose that the variables A, B, and C are all binary (0, 1), with zero the default value. Consider the structural equation model (B1) given by the Boolean equations $A = 1$, $B = 0$, and $C = A \vee B$. Its corresponding graph is pictured in Figure 4.1.

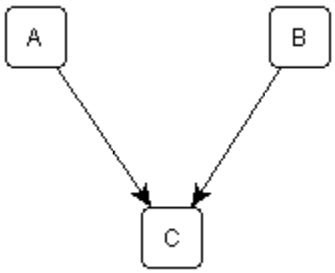

Figure 4.1: Causal Graph for Model B1

The causal network $\boldsymbol{N}_{AC} = \{A, C\}$ is self-contained, since C would equal zero (its default) if A were set equal to zero (its default) and B retained its actual value zero. However, the causal network $\boldsymbol{N}_{BC} = \{B, C\}$ is not self-contained, since C would still equal one if B were set to zero (its default) and A retained its actual value of one.

Using the same variables and the same graph, consider the structural equation model (B2) given by the Boolean equations $A = 1$, $B = 1$, and $C = A \wedge B$. In the model (B2), both networks $\boldsymbol{N}_{AC}$ and $\boldsymbol{N}_{BC}$ are self-contained, since C would equal zero (its default) if either A or B were set equal to zero (their default values).

With the notions of causal networks and default values of a variable in hand, Hitchcock produces TC, which specifies necessary and sufficient conditions for saying that $X = x$ is an actual (or token) cause of $Y = y$.

> TC: Let $<\boldsymbol{V}, \boldsymbol{E}>$ be a causal model, let $X, Y \in \boldsymbol{V}$, and let $X = x$ and $Y = y$. If the causal network connecting X to Y in $<\boldsymbol{V}, \boldsymbol{E}>$ is self-contained, then $X = x$ is a token cause of $Y = y$ in $<\boldsymbol{V}, \boldsymbol{E}>$ if and only if Y counterfactually depends upon X in $<\boldsymbol{V}, \boldsymbol{E}>$. (511)

In other words, if the causal network connecting X to Y is self-contained, then $X = x$ is an actual cause of $Y = y$ if and only if for some non-actual value $x^* \neq x$ of variable X, variable Y takes on some non-actual value $y^* \neq y$.

4.2.3 Fitting TC to Its Work

In Section 4.1, we saw that Hitchcock's theory of actual causation (TC) is intended to be a theory of the folk-attributive concept of causation and that Hitchcock is committed to the FAD with respect to his theory. Now that we have that theory in hand, we can ask how it is supposed to be connected up to folk causal attributions. Perhaps because he does not explicitly say that TC is a theory of the folk-attributive concept of causation, Hitchcock does not say precisely how TC is supposed to work with respect to folk causal attributions.

On my reading of Hitchcock, TC is intended to output some causal attributions when given some specified case(s) as input, and TC is evaluated (at least in part) based on how well its outputs match folk causal attributions about the same case(s). Insofar as TC yields the same causal attributions as do the folk, TC satisfies the FAD and counts as a useful instrument. With respect to the folk, specifying a case might be as simple as telling a story. Folk causal attributions are then whatever the folk happen to say about what causes what in the story. With respect to TC, specifying a case involves specifying a causal structure over a collection of variables and specifying the default (and deviant) values of those variables.

Now, Hitchcock allows both scientific and folk theories to guide the choice of default values for a causal model. But what does that choice amount to? I suggest that there are two possibilities. Either TC is an *instrumental* theory of the folk-attributive concept of causation, or TC is a *realist* theory of the folk-attributive concept of causation. If TC is an instrumental theory, then the theoretician or the experimentalist decides how to model a given story, chooses the default values for the variables in the model, identifies the self-contained causal networks, and decides whether the right sort of counterfactual dependence obtains.[59] If TC is an instrumental theory, then even if it satisfies the FAD, it is uninformative about *how* people come to make the causal attributions they make. Specifically, an instrumental interpretation of TC says nothing about whether people think in terms of self-contained networks, default values, or counterfactual dependence. The instrumental interpretation carries with it no commitments about cognitive architecture.

By contrast, if TC is a realist theory of the folk-attributive concept of causation, then TC does carry with it commitments about cognitive architecture, though these commitments are vague. Instead of a theoretician or experimentalist deciding how to model a given story, choosing the default values for the variables in the model, identifying the self-contained causal networks, and deciding whether the right sort of counterfactual dependence obtains, these tasks are (presumably) left to specific neural mechanisms. Thus, on a realist interpretation, TC not only predicts what ordinary people will say about actual causation for a given story (provided we can make good guesses about the relevant brain processes), it also explains how people come to make the causal attributions they actually make. Hitchcock does not explicitly propose any

[59] These choices need to be made in a principled way in order to avoid satisfying the FAD trivially by clever case-by-case modeling and default-value choices. Otherwise, TC loses its potential usefulness and its ability to genuinely predict what people will say about novel cases.

cognitive models. He does not point to any brain regions as implementing a self-contained-network checker or a default/deviant-value generator.

My guess is that Hitchcock intended to produce an instrumental theory of the folk-attributive concept of causation. In the next two sections, I describe several experiments that I conducted (along with Justin Sytsma and David Rose) in order to test TC under an instrumental reading. If Hitchcock intended to produce a realist theory of the folk-attributive concept of causation, then considerably greater detail is required before TC could be tested. Henceforth, when I talk about TC satisfying the FAD or according with ordinary causal attributions (or failing to do either), I am assuming an instrumental reading of TC.

4.3 DOES TC ACCORD WITH FOLK CAUSAL ATTRIBUTIONS?

In this section and the next, I present experimental evidence that TC does not accord with folk causal attributions. In our experiments, Justin Sytsma, David Rose, and I solicited untutored judgments about some simple examples in which the value of one variable in a self-contained network counterfactually depends on the value of another variable in the network. In these cases, TC counts some action as an actual cause of an event, but ordinary people do not. Surprisingly, untutored participants in our experiments did not maintain that counterfactual dependence of *e* on *c* in a self-contained network is sufficient to count *c* as an actual cause of *e*. Thus, TC does not satisfy the FAD.

Are our examples deviously complex? No. We consider simple variations on an example from the literature on the role of moral considerations in folk causal judgments: a variation on Knobe's (2006) thought experiment about two people logging into an unstable

computer system that subsequently crashes. Not only are these cases simple, they are very similar to cases explicitly discussed by Hitchcock in articulating the consequences of TC.

4.3.1 Study 1: The Lauren and Jane Case

The first example we consider is a variation on Knobe's (2006) Lauren and Jane thought experiment.[60] We produced two variants on the Lauren and Jane Case. The first variant closely mimics the original and tests what causal attributions untutored individuals actually make about cases in which one agent acts permissibly while the other acts impermissibly:

> Lauren and Jane both work for a company that uses a mainframe that can be accessed from terminals on different floors of its building. The mainframe has recently become unstable, so that if more than one person is logged in at the same time, the system crashes. Therefore, the company has instituted a temporary policy restricting the use of terminals so that two terminals are not used at the same time until the mainframe is repaired. The policy prohibits logging in to the mainframe from the terminal on any floor except the ground floor.
>
> One day, Lauren logged in to the mainframe on the authorized terminal on the ground floor at the exact same time that Jane logged in to the mainframe on the unauthorized terminal on the second floor. Lauren and Jane were both unaware that the other was logging in. Sure enough, the system crashed.

The second variant is like the first except that it does not include permissibility information.

In both variants of the Lauren and Jane Case, we assume that the default value for Lauren's action is "does not log in," the default value for Jane's action is "does not log in," and the default value for the computer's state is "not-crashed." Not logging in is a self-sustaining absence. While the value "not-crashed" for the mainframe is not an absence, it is self-sustaining in the sense that if a computer is running, we generally expect it to continue running unless

[60] Interestingly, despite the fact that this example has been identified as an empirical study in the literature (see Roxborough and Cumby (2009, 207) and Hitchcock (2007b, 512) for examples), Knobe did not actually solicit folk reactions to his story. Rather, he presents the Lauren and Jane vignette as a thought experiment. He then asserts that we would attribute the crash more to Jane's behavior than to Lauren's and immediately proceeds to discuss how best to account for "this difference in people's attributions" (68). Thus, it is an open empirical question whether there actually is a difference in people's causal attributions with regard to this case.

something disrupts it. Treating the computer in this way is similar to taking "alive" as the default state for an ordinary human.

Accepting our assumptions about the defaults, the causal network for both variations is self-contained. In fact, both variants have the same form as the causal network (B2) described in Section 4.2. The actual state of the computer at the end of the story—namely, "crashed"—depends counterfactually on the actions of Lauren and Jane. Hence, TC predicts that both Lauren's action and Jane's action will be called actual causes of the computer's crash. However, ordinary people do not call either action an actual cause of the computer's crash.

Participants were randomly assigned one of the two variations. Responses for the Lauren and Jane Case were collected from 195 participants through the Philosophical Personality website (http://www.philosophicalpersonality.com). We excluded 38 participants because they were under 18 years of age, had taken the survey previously, or did not complete the demographic information portion of the survey. Additionally, for the sake of comparison, we excluded 14 participants because they were non-native English speakers or because they had more than minimal training in philosophy. The remaining 143 participants were 73.4% female, with an average age of 34.8 years, and ranging in age from 18 to 81 years old.

Having seen one or the other of the two vignettes, participants were asked: On a scale of 1-7, one being totally disagree and seven being totally agree, how much do you agree with each of the following claims?

1. Lauren caused the system to crash.
2. Jane caused the system to crash.

We found that when permissibility information was included, participants treated Lauren and Jane differently, tending to say that Jane, but not Lauren, caused the system to crash.[61] When permissibility information was not included, participants treated Lauren and Jane identically. Surprisingly, however, they tended to say that *neither Lauren nor Jane* caused the system to crash.[62] The results are shown in Figure 4.2.

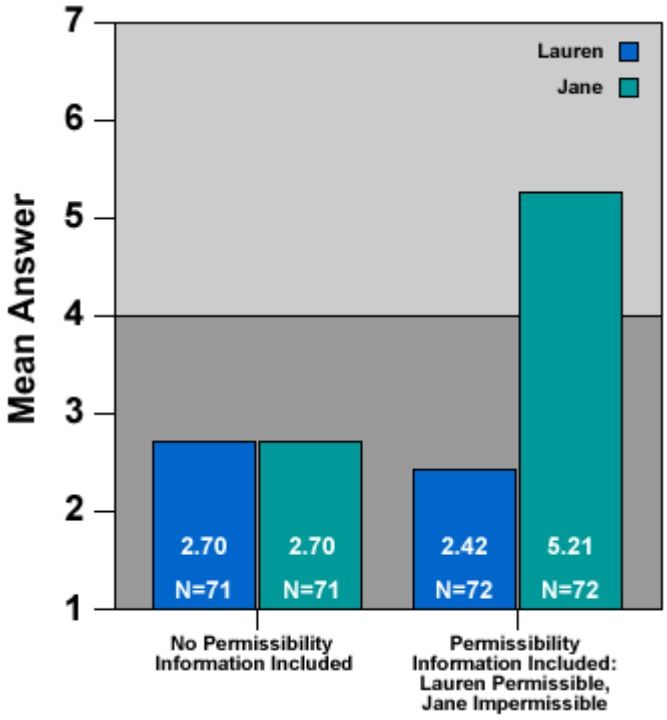

Figure 4.2: Results for Study 1

Recall that TC predicts that ordinary people will say that Lauren's action and Jane's action are *both* actual causes of the computer's crash. When permissibility information is included, ordinary people are only willing to say that one of the two caused the crash. Moreover, in the absence of permissibility information people generally say that *neither* Lauren nor Jane caused

[61] We conducted a paired t-test comparing the mean response for Lauren (mean=2.42, sd=2.04) to the mean response for Jane (mean=5.21, sd=2.19) with the following result: N=72, t(71)=-7.37, p=2.47e^{-10}.

[62] Participants individually said *exactly* the same things about Lauren that they said about Jane in the case without permissibility information. Thus, a t-test comparing the two is not applicable, and t-tests comparing the mean response for each character to the neutral value of 4 are identical. We conducted a one-sided t-test comparing the mean response for Lauren to the neutral value of 4 with the following result: N=71, mean=2.70, sd=2.14, t(70)=-5.10, p=1.387e^{-6}.

the system to crash. Hence, ordinary causal attributions do not satisfy TC, and TC fails to satisfy the FAD.

4.3.2 Study 2: The Action-Centered Lauren and Jane Case

In order to test whether our result was due to an emphasis on agents, as opposed to actions or events, we produced two further variants of the Lauren and Jane Case in which we attempted to focus participants' attention on Lauren and Jane's action of logging in. We call these variants *Action-Centered Lauren and Jane Case*. In one of these two variants, we included permissibility information, and in the other, we did not. The vignettes and scales for these cases were the same as those used in the Lauren and Jane Case (Study 1). However, we changed the statement to be evaluated. Instead of asking how much participants agreed with the claim that Lauren (or Jane) caused the system to crash, we asked them to indicate how much they agreed with each of the following claims:

1. Lauren's action of logging into the terminal caused the system to crash.
2. Jane's action of logging into the terminal caused the system to crash.

We assume that, for both variants of the Action-Centered Lauren and Jane Case, the default value for Lauren is "does not log in" and the default value for the system is "not-crashed." Thus, TC predicts that both Lauren and Jane will be said to have caused the computer to crash.

Responses for the probe *with permissibility information included* were collected from 67 participants through the Philosophical Personality website. We excluded 16 participants because they were under 18 years of age, had taken the survey previously, or did not complete the demographic information portion of the survey. Additionally, for the sake of comparison, 3 participants were excluded because they were non-native English speakers or because they had more than minimal training in philosophy. The remaining 48 participants were 66.7% female, with an average age of 32.8 years, and ranging in age from 18 to 59 years old.

Responses for the probe *without permissibility information included* were collected from 65 participants through the Philosophical Personality website. We excluded 18 participants because they were under 18 years of age, had taken the survey previously, or did not complete the demographic information portion of the survey. Additionally, for the sake of comparison, 4 participants were excluded because they were non-native English speakers or because they had more than minimal training in philosophy. The remaining 43 participants were 74.4% female, with an average age of 30.0 years, and ranging in age from 18 to 63 years old.

Consistent with our previous studies, when permissibility information was included, people asserted that Jane's action but not Lauren's caused the system to crash. When permissibility information was not included, they denied that either Lauren's action or Jane's action caused the system to crash.[63] The results are shown in Figure 4.3.

[63] For the Action-Centered Lauren and Jane Case without permissibility information, we conducted a one-sided t-test comparing the mean response for Lauren's action (mean=3.33, sd=2.31) to the neutral value of 4 with the following result: N=43, t(42)=-1.92, p=0.031. The identical result obtains for Jane. For the Action-Centered Lauren and Jane Case including permissibility information, we conducted a paired t-test comparing the mean response for Lauren's action (mean=2.52, sd=2.06) to the mean response for Jane's action (mean=5.56, sd=2.08) with the following result: t=-6.77, df=47, $p=1.85e^{-8}$.

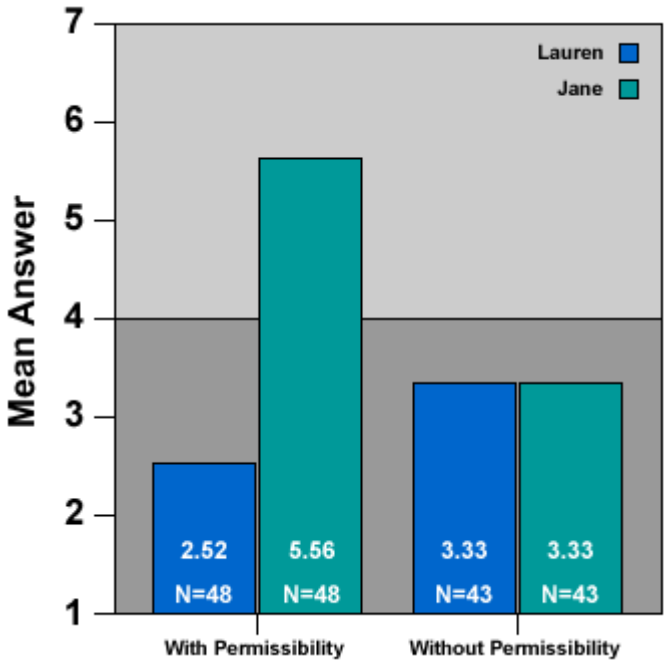

Figure 4.3: Results for Study 2

The results of Study 2 further demonstrate that TC fails to accord with folk causal attributions (assuming that the default value for Lauren is "does not log in" and the default value for the system is "not-crashed"). Though the mean responses for Lauren's action and for Jane's action in the Action-Centered Lauren and Jane Case without permissibility information were slightly higher than the mean responses for Lauren and Jane when the prompt referred to the agents, the difference is not statistically significant at the 0.05 level. More importantly, the mean responses for both *are* significantly less than the neutral value of 4, indicating that untutored individuals deny both that Lauren's action caused the crash and that Jane's action caused the crash.

4.4 FURTHER EVIDENCE

At this point, I think we have adequately shown that TC fails to capture ordinary causal attributions. However, I am not yet prepared to say how often TC differs from ordinary causal attributions or why it does so. Some further results both shore up the case against the descriptive

adequacy of TC and also exclude some explanations for the causal attributions we observed in the previous two studies. I consider the further results now.

4.4.1 Study 3: The Lauren Alone Case

Perhaps participants were reluctant to say that either of two agents caused an outcome if neither agent's action was sufficient for that outcome and nothing further distinguishes the two. In order to test this possibility, we simplified the Lauren and Jane vignette (with no permissibility information) by removing Jane entirely. The modified probe reads as follows:

> Lauren works for a company that uses a mainframe that can be accessed from terminals in its building. Though the company does not know it, the mainframe has recently become unstable, so that if anyone logs into the system, the system crashes.
>
> One day, Lauren logged into the mainframe. Sure enough, the system crashed.
>
> On a scale of 1-7, one being totally disagree and seven being totally agree, how much do you agree with the following claim?
>
> Lauren caused the system to crash.

We assume that in this case, the default value for Lauren is "does not log in" and the default value for the system is "not-crashed." Thus, TC predicts that ordinary people will say that Lauren caused the system to crash.

Responses for this probe were collected from 86 participants through the Philosophical Personality website. We excluded 21 participants because they were under 18 years of age, had taken the survey previously, or did not complete the demographic information portion of the survey. Additionally, for the sake of comparison, 4 participants were excluded because they were non-native English speakers or because they had more than minimal training in philosophy. The remaining 61 participants were 73.8% female, with an average age of 31.5 years, and ranging in age from 18 to 69 years old.

As in Study 1, people denied that Lauren caused the system to crash.[64] The results are shown in Figure 4.4.

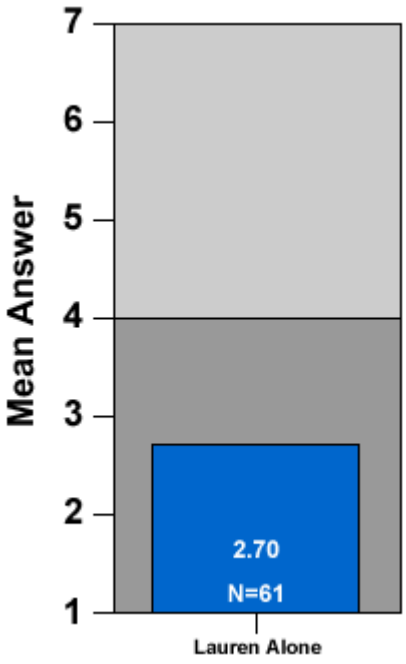

Figure 4.4: Results for Study 3

Once again, our results indicate that given some plausible assumptions about the default values of the variables in the model, TC fails to accord with ordinary causal attributions.

4.4.2 Study 4: The Lauren and Jane Physically Caused Case

Since participants treat the agents differently depending on whether they act permissibly or impermissibly (as well as for other reasons not discussed in the present paper), we suspect that ordinary usage of "cause" has broadly moral or prescriptive connotations. Hence, we wondered what participants would say if we asked whether Lauren or Jane had *physically* caused the system to crash. Somewhat surprisingly, the results were very similar to those reported for the first Lauren and Jane Study. When given permissibility information, participants indicated that

[64] We conducted a one-sided t-test comparing the mean response for Lauren (mean=2.705, sd=2.02) to the neutral value of 4 with the following result: N=61, t(60)=-5.01, p=2.56e^{-6}.

Jane but not Lauren physically caused the system to crash. Without permissibility information, participants denied that either one caused the system to crash.

Responses were collected from 192 participants through the Philosophical Personality website. Participants were randomly assigned to a vignette with or without permissibility information included. We excluded 52 participants because they were under 18 years of age, had taken the survey previously, or did not complete the demographic information portion of the survey. Additionally, for the sake of comparison, 13 participants were excluded because they were non-native English speakers or because they had more than minimal training in philosophy. The remaining 127 participants were 74.0% female, with an average age of 32.7 years, and ranging in age from 18 to 71 years old.

As in Study 1, when given permissibility information, people indicated that Jane but not Lauren caused the system to crash, and when given no permissibility information, people denied both that Lauren caused the system to crash and that Jane caused the system to crash.[65] The results are shown in Figure 4.5.

[65] For the probe with permissibility information included, we conducted one-sided t-tests comparing the mean response for Lauren (mean=2.02, sd=1.66) to the neutral value of 4, with the result that N=64, $t(63)=4.24$, $p=3.78e^{-5}$ and comparing the mean response for Jane (mean=5.15, sd=2.18) to the neutral value of 4, with the result that N=64, $t(63)=-9.58$, $p=3.25e^{-14}$. The responses for Lauren and for Jane differ significantly, according to a paired, two-sided t-test: N=64, $t(63)=-8.47$, $p=5.29e^{-12}$. For the probe with no permissibility information included, we conducted one-sided t-tests comparing the mean response for Lauren (mean=2.51, sd=2.06) to the neutral value of 4, with the result that N=63, $t(62)=-5.74$, $p=1.51e^{-7}$ and comparing the mean response for Jane (mean=2.32, sd=1.91) to the neutral value of 4, with the result that N=63, $t(62)=-7.00$, $p=1.07e^{-9}$. The responses for Lauren and for Jane did not differ significantly, according to a paired, two-sided t-test: N=63, $t(62)=1.426$, $p=0.159$.

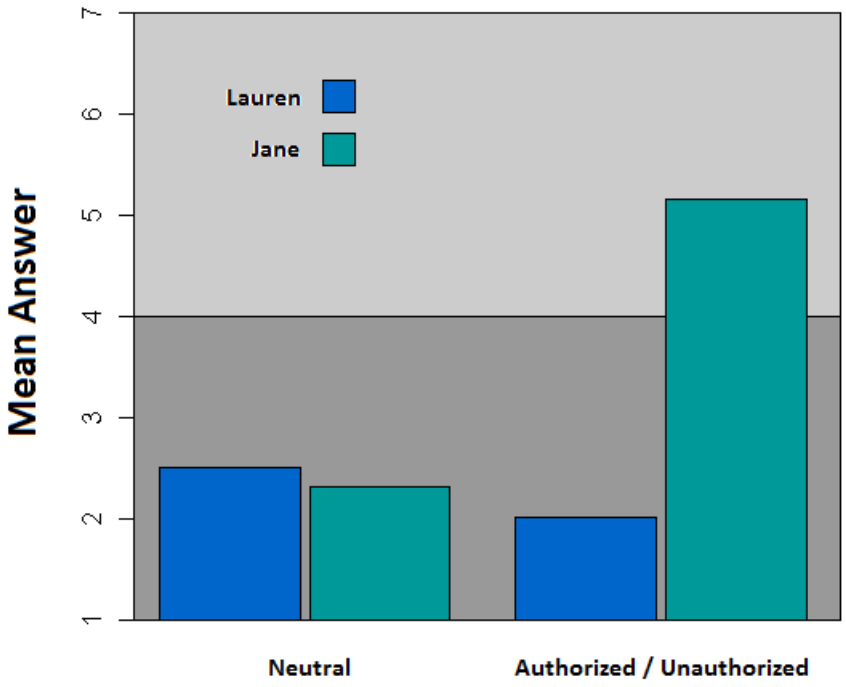

Figure 4.5: Results for Study 4

Once again, our results indicate that given some plausible assumptions about the default values of the variables in the model, TC fails to accord with ordinary causal attributions.

4.4.3 Study 5: The Multiple-Choice Lauren and Jane Case

We again presented participants with the vignettes from Study 1; however, instead of asking how much they agreed with some claim or other, we asked the multiple-choice question, "Who do you think caused the system to crash?" For which we provided the following options: (1) Lauren, (2) Jane, (3) Both Lauren and Jane, (4) Neither Lauren nor Jane, (5) Other. And we asked those who answered "Other" to briefly tell us why.

Responses were collected from 128 participants through the Philosophical Personality website. Participants were randomly assigned to a vignette with or without permissibility information included. We excluded 28 participants because they were under 18 years of age, had taken the survey previously, or did not complete the demographic information portion of the

survey. Additionally, for the sake of comparison, 7 participants were excluded because they were non-native English speakers or because they had more than minimal training in philosophy. The remaining 93 participants were 63.4% female, with an average age of 34.8 years, and ranging in age from 18 to 79 years old.

When given permissibility information, people slightly preferred to say that Jane caused the system to crash than to say that both had caused the system to crash. When given no permissibility information, people very much preferred to say that neither caused the system to crash than to say that both had caused the crash. (Only one person indicated that Lauren caused the crash and no one indicated that Jane caused the crash.) The results are shown in Figure 5.6.

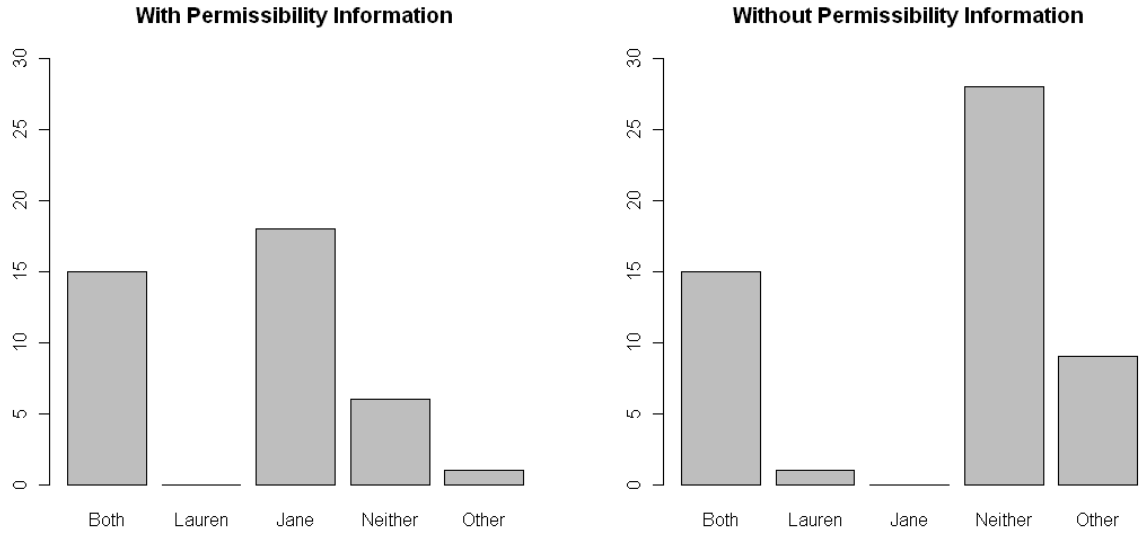

Figure 4.6: Results for Study 5

If TC is descriptive of ordinary causal attributions, the modal answer should be "Both." But the modal answer is "Jane" when permissibility information is given and "Neither" when it is not.

The explanations from participants who answered "Other" in the no-permissibility-information condition were interesting but also not encouraging for TC. Most of the responses (8 out of 9) suggested that the crash was "the fault" of the company, some other employees of

the company (e.g. the IT department), or the computer system itself. Taking up the last of these suggestions, one might model the scenario with four variables instead of three: two variables for the computer system (one for its state before the log-in and one for after) and one variable each for Lauren and Jane. Let C_0 and C_1 be variables—with possible values of stable (default), unstable, and crashed—representing the state of the system at the beginning of the work day and after the log-in, respectively. Then the causal graph is given in Figure 4.7:

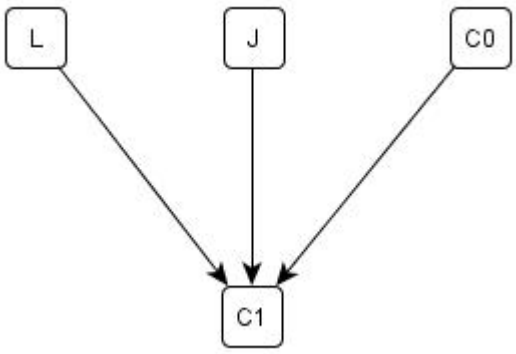

Figure 4.7: Modified Causal Graph for the Lauren and Jane Case

The actual values of the variables are J = logs in, L = logs in, C_0 = unstable, and C_1 = crashed. The substantive equation corresponding to the graph is such that the value of C_1 is "crashed" if the value of J is "logs in" and the value of L is "logs in," and the value C_1 is the same as the value of C_0 otherwise.[66] Each of the causal networks $\{J, C_1\}$, $\{L, C_1\}$, and $\{C_0, C_1\}$ in the model is self-contained, and the value of C_1 depends counterfactually on the value of all three other variables in the model. Hence, according to TC, the values of all three variables should count as actual causes of the value of variable C_1. But that does not fit what ordinary people say. So, once again our results indicate that given some plausible assumptions about the default values of the variables in the model, TC fails to accord with ordinary causal attributions.

[66] If we let "logs in" be 1 and "does not log in" be 0 and if we let "crashed" be 0, "unstable" be 1, and "stable" be 2, then the equation $C_1 = (1 - J \cdot L) \cdot C_0$ correctly describes the system.

4.5 SOMETHING MUST GO

Something has to go. I claim that what ought to go depends on one's goal. If the goal is to predict what people will say about causation, then the background theory needs to be radically altered. By extension, if the theory is good at predicting what people will say about causation, then the theory will be useful for the purposes to which people put causal judgments. A growing body of literature indicates that when people make causal judgments they usually intend to be attributing responsibility or to be attributing praise or blame to agents. If the goal is to do good diagnostic work, then the commitment to correctly capture what ordinary people say about causation ought to be dropped. With diagnosis as the goal, the theory transforms from a descriptive theory into a normative theory. In this respect, a theory of actual causation is a metaphysical theory, and Hitchcock (2007b) is wrong to say that the metaphysical concept of causation is a dispensable hybrid of the scientific and folk-attributive concepts.

I claim that the main goal for a theory of actual causation is facilitating diagnostic success. One wants to identify the actual causes of some outcome in order to (a) prevent that outcome from continuing to happen or (b) succeed in producing that outcome on demand. Hence, the way to judge whether a theory of actual causation is adequate is by how well it allows one to prevent unwanted outcomes from continuing to occur and to bring about desirable outcomes on demand. These two modes of actual causation correspond to dependence (necessary conditions) and production (sufficient conditions), respectively.

As pointed out in Section 4.3.2, actual causation is routinely connected to historical explanations and to attributions of moral and legal responsibility.[67] In fact, actual causation is

[67] Historians often explain (or attempt to explain) why events happened as they did, i.e. we want to know why the American civil war happened or why the Soviet Union fell apart. Many ethicists and legal theorists hold agents

central to explanation and to fault attribution more generally as well. Actual causation is the conceptual underpinning of diagnostics, of which historical explanation and responsibility attribution are instances. Providing a diagnosis of some state of affairs is to identify the actual cause(s) of that state of affairs.

Diagnostic problems are everywhere. They range from medical diagnosis and criminal forensics to computer technical support and automotive repair. A diagnostician wants to determine the underlying actual cause of a constellation of observed symptoms. Usually, though not always, the symptoms are regarded as problematic and the goal is to eliminate them, i.e. the diagnostician wants to identify why a system is broken. A diagnosis might be guided by knowing what generally causes the observed symptoms, but the diagnostician will not be content to learn that in *most* cases the observed symptoms have a given cause. Rather, the diagnostician wants to know why *this unit* exhibits the observed symptoms (at least up to equivalence of treatment).

Hitchcock and Knobe (forthcoming) make an interesting suggestion along these lines, though they never explicitly connect actual causation with diagnostics. Instead, they claim that out of all the variables in a causal structure that *could* be manipulated to achieve some desired effect, the actual causes are those variables that *should* be manipulated in order to achieve the effect. Thus, according to Hitchcock and Knobe, "[T]he concept of actual causation enables us to pick out appropriate targets for intervention." I agree that actual causes often make the best targets for interventions; however, I have two related worries. First, I do not know how often the

---

morally and legally responsible for bad things that happen in the world in part because those bad things are thought to be effects of actions (or inactions) of the agents. For example, if Smith is part of a mob that beats Jones to death, then Smith is legally responsible for Jones' death in part because Smith was an actual cause of Jones' death. One sometimes encounters cases where a person is judged to be an actual cause of some undesirable occurrence and yet is not held responsible (owing to intent, foreseeability, etc.), but converse cases are rare if they exist at all. However, see Sartorio 2004 for a much more nuanced discussion.

best targets for interventions are also the actual causes. Sometimes, they are not the same. For example, suppose a system has a large, heavy part A and a small, light part B. Further suppose that part A moves out of alignment and the system stops functioning. As it turns out, the system could be repaired in two ways: (1) move part A back to its original position or (2) move part B to a new position. The position of part B may be the better target of intervention—if moving part B is simpler, more efficient, more economical, etc.—and yet, I have no inclination to say that part B remaining in its original position was an actual cause of the system breaking. Second, while I agree that norms are important for judgments of actual causation, I think that the relevant norms apply to systems taken as wholes, as opposed to parts of systems. Typically, what sets the default states for a system (provided it has defaults) is an idealization of the system as a whole, not just an idealization of its output. A blueprint is an example of the relevant sort of idealization. I suspect that these judgments will be shared by ordinary people, though that suspicion should be checked empirically.

In order to make diagnostic judgments, one needs to know not just how the variable values in the system depend on one another. One needs to know how the system under study is supposed to work. An ideal to which a system is compared need not be a fact about the world. Rather, an ideal may represent a stance that one takes towards a system when evaluating it relative to some purpose. For example, one might deny that the heart was designed and yet adopt a design stance towards the heart in diagnosing a heart illness. To call a heart illness an illness is already to attribute some proper functioning to the heart—to compare it to an ideal or norm. But the standard or norm need not be a feature of the world; it need not be an objective fact. If appeal to norms is essential to actual causation judgments, then purely formal accounts of actual causation are doomed to failure, since they cannot incorporate context-specific features.

Indeed, one might formally flag some variable values as defaults, but doing so will not help one to apply the theory—to actually attach the default label to the correct variable values from one case to the next.

CHAPTER FIVE

META-THEORY

In Chapter 3, I described several competing theories of actual causation and appealed to my intuitions (and the intuitions of the reader) in criticizing those theories. In Chapter 4, I discussed some experiments designed to discover what causal attributions ordinary people make in some simple cases. An account of actual causation developed through experimentation as in Chapter 4 would contribute to a naturalistic, descriptive project. Such an account (if successfully developed) would tell us how people in fact reason about actual causation.

What about the prospects for a normative account of actual causation? In the present chapter, I step back and ask how we ought to decide which one of several competing normative theories of actual causation (or of any other basically metaphysical concept) is correct. I consider two basic approaches to meta-theory: Socratic and Euclidean.

5.1 TWO APPROACHES TO META-THEORY

A man is shot and subsequently dies. The medical examiner reports that the immediate cause of death was exsanguination and that the proximate (or underlying) cause was a gunshot wound to the chest. Determining that these were the immediate and proximate causes of death seems pretty straightforward. But *what* makes exsanguination a cause in this case? And *what* makes the gunshot or the gunshot wound a cause? In virtue of what do these things count as causes?

Many substantive theories of causation attempt to say what, *in general*, makes something a cause.[68] Similarly, as we saw in Chapter 3, several different theories attempt to say what is special about *actual* causation.[69] But there is a prior question that I would like to consider. *How* are such metaphysical questions to be answered? Given a solution to the metaphysical puzzles, we can often carry out experiments that provide evidence for this or that particular causal claim. For example, we want to know whether the gunshot caused the man to die. If we thought that actual causation just amounted to difference-making, we could take a gelatin dummy to a firing range, reproduce as closely as possible the circumstances of the shooting, and see how much difference a bullet makes. We could consider other people similar to the gunshot victim, except for the shot itself and see what difference the gunshot makes. We might need to do significant work to figure out the details of the causal structure (causal laws) governing variables like bullet type, bullet velocity, location of injury, severity of injury, etc. But after building a suitable model of the causal system, we would have an empirically contentful answer to the question: did the gunshot cause the man to die?

Given an answer to the metaphysical problem, we could conduct experiments to figure out what caused what in a specific circumstance. But could we conduct experiments to figure out which *metaphysical theory of causation* is correct? I think the answer is a qualified yes. Given some plausible bridging principles, experiments might serve to constrain metaphysical theorizing about causation. These bridging principles are actually quite general. They belong to very basic philosophical frameworks, and they apply to lots of problems, not just to causation.

---

[68] Causal theorists have offered regularity accounts (Mackie 1965, 1974; Beebee 2006; Baumgartner 2008; Psillos 2009), probabilistic accounts (Suppes 1970; Cartwright 1979; Eells 1991; Hitchcock 1993; Williamson 2009), counterfactual accounts (Lewis 1973, 2000; Paul 2009), process accounts (Salmon 1984, 1994, 1997; Dowe 1992, 2000, 2009), mechanism accounts (Machamer et al. 2000; Craver 2007; Glennan 1996, 2009), agency and manipulability accounts (Dummett and Flew 1954; Hausman 1986; Menzies and Price 1993; Woodward 2003, 2009), and projectivist accounts (Blackburn 1984; Price 1991, 2007; Beebee 2007).

[69] I will not consider here whether taking actual causation seriously requires a commitment to some kind of pluralism. For an overview of pluralism about causation, see Hitchcock (2007) and Psillos (2010).

Given their generality, I will set out the frameworks first and then apply them to the specific problem of actual causation.

The basic contrast has been described in different ways by different writers. I am especially fond of Glymour's characterization of the two frameworks. In a review of Woodward (2003), Glymour (2004) contrasts two approaches to philosophical theorizing: Socratic and Euclidean. The Socratic approach is characterized by a search for necessary and sufficient conditions for the application of a concept or term. Many early and middle Platonic dialogues have this form. Socrates meets someone at a party or at the courthouse or at the gym. The unfortunate soul makes a confident assertion about some abstract entity or property, like beauty or piety or friendship. "You say that such and so is pious," Socrates says to his interlocutor (victim). "Then you must know what piety is. Please tell me, so that I will be wise!" The interlocutor offers a list of things that exemplify the abstraction, to which Socrates replies, "I don't want a list; I want an account of the essence of the thing." The interlocutor proceeds to offer theory after theory, as Socrates presents counter-examples to each one in turn until the conversation ends—usually because the interlocutor has to go somewhere in a hurry.

The Socratic approach is closely related to conceptual analysis, and it is conducted primarily by the method of cases, wherein ordinary (or not-so-ordinary) judgments, called *intuitions*, are the standard currency. For example, I tell you a story. Joe has an important meeting in the morning, so he sets two alarms—a plug-in digital alarm clock and a hand-wound analog alarm clock—to go off at precisely 6:35 AM. Suppose that the two alarms are redundant in the sense that Joe will wake up if either alarm goes off. As it happens, both alarms go off at exactly the same time, and Joe wakes up. If you have the unsupported snap judgment, or intuition, that both alarms caused Joe to wake up, then you might think that there is something

objectionable about simple difference-making theories of causal dependence like that in Lewis (1973). If you think that each alarm, by itself, caused Joe to wake up, then you might think that there is something objectionable about collectivist accounts of causal power. Getting our intuitions right (with some possible debate over the scope of "our") is supposed to be the goal in the Socratic approach to philosophy.

The Euclidean approach is more closely related to explication. Instead of attempting to satisfy ordinary intuitions about cases or else to explain them away, one picks out an interesting feature or set of features of some philosophically interesting concept. One then lays down some axioms as a good-enough approximate formalization of the interesting features, and then one derives consequences from the axioms. Some failures of intuitiveness are expected, since the goal is not to fit intuitions. Glymour (2004) argues that since Euclidean theories are not burdened with the task of fitting intuitions, Euclidean theories "have the virtue that their consequences can be interesting, informative and non-obvious from their starting assumptions" (780). Euclidean theories are typically inaccurate in one way or another. As Protagoras objected, Euclidean points, lines, and planes are nowhere to be found in the world of real experience. However, many things are enough like points, lines, and planes that the Euclidean axioms (and the theorems derived from them) are immensely useful.

If adopted as ideologies, the Socratic and Euclidean approaches are antagonistic. However, we need not demand ideological purity, and if we do not, then the two approaches are potentially complementary. Axiom systems implicitly define their primitive terms by indicating how those primitive terms are to be used. Hence, one might endorse an axiomatic system as an analysis of the primitive terms and then defend that analysis via Socratic examples and counter-examples. Similarly, every analysis (or at least the ones that are clear enough), could be set

down in terms of axioms. Lewis' various theories of causation could be given axiomatically. Actually, doing so would not be too difficult, since Lewis himself produced axiom systems for various counterfactual logics. Pearl (2000, Chapter 7) shows how to inter-translate structural equation approaches and possible-worlds approaches to counterfactuals. Since Lewis derives his theory of causation from his account of counterfactuals, it is only a short step from this to inter-translating theories of causation. It seems then that the only reason Socratic theories do not have interesting, non-obvious consequences is that either no one bothers to look for them or else when they are found, they are treated as evidence against the correctness of the theories.[70] Moreover, Euclidean theories have much to gain from good Socratic work. Until enough Socratic work has been done, axiom systems are not likely to be very fruitful, since they are not likely to have a close enough connection with any interesting feature of the world.

5.2 THE SOCRATIC FRAMEWORK

The main theme of the Socratic approach to philosophical theorizing is fitting intuitions. But within the Socratic framework, disagreement exists about the nature of intuitions and the process whereby one fits a theory to some intuitions. The two major positions are Cartesian rationalism and deflationary rationalism. The two positions share the core commitment to adjudicating theoretical disputes by appeal to intuitions. But they disagree about the nature of intuitions and (to some extent) about the process by which disagreements are settled. The possible connections to experimentation are similar in both cases.

[70] Ichikawa (2009, 114) calls for greater boldness on the part of conceptual analysts: "There are counterintuitive truths; finding one needn't be cause for embarrassment. In cases where counterintuitive consequences of otherwise appealing theories are discovered, my advice to philosophers is to be upfront. Gild the pill with an explaining-away if you have a plausible one to offer; if you don't, then admit that you have a counterintuitive consequence to swallow, and explain why it's worth it to do so. Weak attempts to explain away recalcitrant intuitions only further muddy the issue."

5.2.1 Cartesian Rationalism

The classic statement of Cartesian rationalism is, of course, Descartes' claim in his *Discourse on Method* that whatever one clearly and distinctly conceives is true. On this view, all philosophical problems are problems of conceptual analysis. The philosopher spies some interesting topic, *T*, like justice or knowledge. To have a true account of *T* is just to have a clear and distinct conception of *T*. One has a clear and distinct conception of *T* just in case one has a set of necessary and sufficient conditions for the application of the concept or the use of a set of terms. Hence, to have a true account of *T* is to have a conceptual analysis of *T*. For example, an account of knowledge specifies the conditions under which it is correct to say that such and so is an example of knowledge. If an account does that, then it is a true account of knowledge.

In giving an account of some topic *T*, the rationalist deploys a method of cases, which has a venerable history. Since the time of Plato's early Socratic dialogues, philosophers (or at least some of them) have employed the method in basically its present form as follows: (1) collect intuitions with respect to topic *T*—e.g., justice, courage, piety, etc.; (2) formulate a theory that captures these intuitions; (3) look for counter-examples; (4) formulate a new theory.[71] Because counter-examples *inevitably* crop up for philosophically interesting topics, the result of this method is a steady complicating of philosophical theories. We have already seen this sort of thing in Chapter 3 with theories of actual causation. For further examples from the causation literature, consider the history of "the" probabilistic theory of causation or the history of "the" counterfactual theory of causation.

As we will see, the method of cases functions similarly in the Cartesian and naturalized versions of rationalism, for the most part. However, the Cartesian version of rationalism (or

[71] One might say that the form of rationalism at stake here is Platonic, not Cartesian, but the role of intuition is, I think, more clearly articulated by Descartes than by Plato.

intuitionism) endorses the claim that humans have some special intellective faculty that tracks the truth about philosophically interesting concepts. This intuitive faculty produces intuitions, which are by their very nature clear and distinct conceptions. In the hands of a Cartesian rationalist, the goal of the method of cases is to fit theory to fixed intuitional data. That is, once we identify the intuitions, the theory has to be responsible to them, because our intuitions track the truth about philosophically interesting topics.

5.2.2 Deflationary Rationalism

Many philosophers find robust, Cartesian-style intuitions too spooky to countenance. Some of these philosophers want to abandon the entire rationalist enterprise, as we will see below. However, some philosophers think that "intuition" may be given a deflationary sense and still be fit for its work in philosophical discourse. Instead of supposing that intuitions are the products of a special intellectual faculty, intuitions might just be a kind of judgment (Williamson 2007, 2009), an opinion (Lewis 1986), or a fast-and-dirty brain process (Sloman 2002; Kahneman and Frederick 2002; Kahneman 2003; Evans 2003). Or philosophers might just be using "intuitively …" as a simple, colorful substitute for "it seems to me that …" (Dorr 2010).

But after one gives up on the idea that humans have a special intuitive faculty that tracks the truth about philosophically interesting concepts, what justification can be given for the Socratic method of cases? Here is one suggestion. One might say that metaphysical problems are not the sort of problems that have robust, factual solutions. Rather, solutions to such metaphysical problems are like choosing a grammar or a language, or like deciding to play a particular language game. And hence, instead of trying to reach bedrock truth, deflationary rationalists settle for what Rawls (1971) calls reflective equilibrium.

About a hundred years before Rawls, Peirce (1992 [1877]) described something similar to the Socratic method of cases—what Peirce called the *a priori* method of fixing beliefs—as follows:

> Let the action of natural preferences be unimpeded, then, and under their influence let men, conversing together and regarding matters in different lights, gradually develop beliefs in harmony with natural causes. This method resembles that by which conceptions of art have been brought to maturity. The most perfect example of it is to be found in the history of metaphysical philosophy. Systems of this sort have not usually rested upon any observed facts, at least not in any great degree. They have been chiefly adopted because their fundamental propositions seemed "agreeable to reason." This is an apt expression; it does not mean that which agrees with experience, but that which we find ourselves inclined to believe. (118-119)

The central problem for the deflationary rationalist's use of the method of cases, as Peirce notes, is that application of it has not historically produced long-term agreement. That is, the method of cases does not have any record of actually leading to reflective equilibrium. Peirce points out the failure with respect to different ages of philosophical speculation:

> [The method's] failure has been most manifest. It makes of inquiry something similar to the development of taste; but taste, unfortunately, is always more or less a matter of fashion, and accordingly metaphysicians have never come to any fixed agreement, but the pendulum has swung backward and forward … from the earliest times to the latest. (119)

Stich (1998) raises the possibility of different equilibria at the same time but in different places, and some of the most widely discussed experimental work bears on this question of the spatial uniformity of intuitions at a single time (Machery et al. 2004; Sytsma and Livengood forthcoming).

5.2.3 Broad and Narrow Rationalisms

Cartesian and deflationary rationalism both admit of a range of varieties, corresponding to how one characterizes the community of inquirers or to whose intuitions matter for philosophical theorizing. *Broad rationalism* endorses the view that characteristically philosophical concepts, like causation, are shared human concepts that require no special training to master. That is, any

adult human with ordinary faculties will have case-by-case intuitions about causation that are as good as those of any so-called "expert." *Narrow rationalism* endorses the view that even concepts like causation, which appear to be widely shared, are better understood by experts who have reflected on them. Hence, one ought to pay attention to the intuitions of experts or at least give them greater weight when constructing a theory.

One might be tempted to narrow down the relevant community to a single individual. However, construed in that way, one probably has to commit to Cartesian rationalism for the sake of coherence. And even then, the view faces a serious challenge. As Peirce put it:

> To make single individuals absolute judges of truth is most pernicious. The result is that metaphysicians will all agree that metaphysics has reached a pitch of certainty far beyond that of the physical sciences;--only they can agree upon nothing else.

The problem is that if the relevant community is reduced to a single individual, then each individual (who represents a separate community) has no constraints on beliefs and no two individuals have any objective way to adjudicate or arbitrate disputes. Consequently, the rationalist usually will widen the circle a little bit—taking in the experts at least. In any event, rationalists hope that the method of cases eventually halts and that the community—however it is constituted—comes to some agreement about the correct theory of *T*, either because that is the explicit goal of the method (in the deflationary case) or because the method is believed to track the truth about ordinary concepts (in the Cartesian case).

5.2.4 Experimental Rationalism

At first blush, rationalism seems perfectly antithetical to experimentation. How might experimentation be useful for a rationalist? Here are three possible ways in which experiments might be useful for rationalists.

First, even if one endorses the claim that humans have a special intellectual faculty that tracks conceptual truths, one might wonder what sort of access humans have to the outputs of that faculty. Traditionally, Cartesian rationalists have claimed that we have perfect introspective access to our intuitions. However, serious doubts have been raised about the extent of human introspective powers (Gopnik 1993; Peirce 1992; Shoemaker 1988; Schwitzgebel 2004, 2007, 2008). One might believe that humans have an intuitive faculty but deny that humans have the ability to distinguish genuine intuitions from merely apparent ones by introspection. For example, Ludwig (2007) takes an intuition to be "an occurrent judgment formed solely on the basis of competence in the concepts involved in response to a question about a scenario" (135); however, he concedes, "It can seem to one that a judgment is an intuition when it is not, just as it can seem to one that one remembers something when one does not" (137). In a footnote, Ludwig suggests that the function of the method of reflective equilibrium is to separate the genuine intuitions from the judgments that merely seem to be intuitions. He writes, "The issue that reflective equilibrium addresses, in these terms, is how to determine what are the real as opposed to seeming intuitions, and it recommends those that fit with most of the others" (137, n. 20).

Although Ludwig is not friendly to experimental methods in philosophical research, his view of intuitions and the method of reflective equilibrium suggests a role for experimentation: helping to distinguish the genuine intuitions from the merely apparent ones. This could be done in a couple of different ways. Insofar as one believes that humans have a special intuitive faculty of mind, one could attempt to identify that faculty by experimentation. Ask subjects questions about conceptual truths while in an fMRI or PET scanner, for example. Alternatively, one might use pencil and paper methods to elicit large collections of "intuitions" and then follow Ludwig's

advice in trying to harmonize them. Such an approach could be used regardless of whether one thinks that the relevant "intuitions" to be harmonized together are the "intuitions" of a single person or those of a community of people.

Second, rationalists like Descartes thought of intuitions as incorrigible and error-free, but many contemporary rationalists think that intuitions vary in quality and may be improved with training (Bealer and Strawson 1992; Bealer 1998, 2002; Sosa 1998; Weatherson 2003). If one thinks that intuitions track the truth imperfectly, then even if one has introspective access to them, one might not want to trust the intuitions of any single individual. Instead, one might take advantage of the power of multiple witnessing in order to discover conceptual truths. Insofar as one thinks that training improves intuitions, one might give different weights to different intuiters based on their level of experience.

Third, one might think that intuitions have varying strengths. Hence, philosophers sometimes say that philosophical theories need to satisfy the *strongest* intuitions, but they suppose that weak intuitions may be safely ignored. However, as with the genuineness of intuitions, one might worry that the strength of an intuition is not reliably accessible via introspection. Perhaps the strength of an intuition may be elicited by betting experiments; whereas, straightforward introspective reports are unreliable. If the strength of an intuition may be determined by experiment but not by introspection, then experiment has a valuable role to play in the rationalist program.

5.3 THE EUCLIDEAN FRAMEWORK

In contrast with the Socratic approach to philosophy, the Euclidean approach is not ultimately concerned with conceptual analysis. Rather, Euclidean philosophical projects often *begin* with a

rough-and-ready conceptual analysis in order to *go on* to see what consequences such an analysis has for conceptual and practical life. If the starting point for the Euclidean does not represent existing concepts very well, that is hardly a problem. Euclidean philosophers may sometimes be concerned with clarifying existing concepts or practices, but more often, they are concerned with reforming existing concepts and practices (Glymour 2004, 779-780). Hence, as Glymour says, Euclidean theories are justified by their fruits. In this respect, at least, the Euclidean approach to philosophy has a very pragmatic spirit. I am not sure whether every Euclidean theorist must be a pragmatist, but pragmatism harmonizes well with the Euclidean approach to philosophical theorizing.[72]

5.3.1 Pragmatism

One way to understand pragmatism as a philosophical school is as a reaction to Cartesian rationalism. Three of Charles Peirce's earliest papers (all published in the *Journal of Speculative Philosophy*) present a concerted attack on rationalism. The other founding father of pragmatism, William James, dedicates his lectures on pragmatism to the memory of John Stuart Mill, perhaps the most vociferous critic of rationalism—or as he called it, "intuitionism"—to ever live. James writes in his dedication that he "likes to picture [Mill] as our leader were he alive to-day." In his autobiography, Mill writes that the whole point of his mammoth *System of Logic* was to expel the intuitionists from their stronghold in logic, mathematics, and the physical sciences. And Mill's lengthy *Examination of Sir William Hamilton's Philosophy* is an extended critique of intuitionism in metaphysics and logic.

[72] As a pragmatist, I tend to see pragmatists, people who lean toward pragmatism, and pragmatic elements in philosophical theories all around me. I recognize that this is a cognitive bias. I hope that my melding of the Euclidean framework and a version of philosophical pragmatism is interesting enough that the reader will forgive my shortcomings here.

Pragmatists differ from rationalists in a number of ways, but in the current context, three stand out as especially important: (1) pragmatists do not think there is any reason to postulate a special intuitive faculty; (2) pragmatists do not think that intuitions—if they exist at all—have any marks that make them introspectably distinct from other mental entities, like memories, perceptions, or conclusions of arguments; and (3) pragmatists think that concepts are exhausted by their practical consequences. Taken together, the pragmatist has a very different proposal for evaluating the truth of philosophical claims. Instead of asking whether a theory or concept of *T* agrees with intuitions about cases, the pragmatist asks what the theory or concept of *T* aims to do and whether it achieves its aim well or ill. (Satisfying its aim optimally then goes proxy for truth, or if you are squeamish about such things, satisfying an aim optimally indicates truthiness.) The pragmatist does not need a method of cases. Rather, the pragmatist only needs a well-defined aim and a standard of satisficing for that aim.

5.3.2 Experimental Pragmatism

Pragmatism is characterized by the acceptance of some variation on the view that concepts are exhausted by their practical consequences. Instead of asking about the nature of a concept or attempting to give a conceptual analysis, a pragmatist asks what the concept does or what its role is. Like other experimental philosophers, experimental pragmatists think that if one wants to know what a concept is good for, one has to go out and look.

Experimental pragmatists have an affinity with so-called natural language philosophers, including the later Wittgenstein, Austin, Strawson, and others. Pragmatists have two kinds of project. First, they conduct experiments to discover the roles that different concepts are meant to play or the purposes that different language games serve. Second, they conduct experiments

designed to uncover what costs and benefits are conferred on individuals or communities that have certain concepts or play certain language games.

In one sense, experimental pragmatists and experimental rationalists are after the same targets: concepts. Rationalists aim to give an analysis of concepts in terms of intuitions about case-by-case applications of the concepts. Pragmatists aim to give an account of concepts in terms of the roles that the concepts play and the benefits they confer on their users. However, the two projects are significantly different. The rationalist project assumes that concepts (or their objects) float free from how any community actually uses them. When a rationalist provides an analysis of the concept "knowledge," for example, that analysis is supposed to tell us something about what knowledge is *really like*. The pragmatist project, on the other hand, supposes that concepts (and their targets) are relative to communities of users or relative to a choice of language game, and hence, they deny that there is any fact of the matter about what knowledge is really like, for example.

Moreover, the pragmatist does not stop after the aim of a concept is identified. The pragmatist project has two components or phases: one descriptive and one prescriptive. In the descriptive phase, the pragmatist tries to discover what a theory or concept is supposed to do—what is its purpose or aim. In the prescriptive phase, the pragmatist tries to evaluate the theory or concept relative to its purpose. In the first phase, at least, experiments might be quite useful. One might, of course, simply state by fiat that some theory or concept aims at such and so. Or one might identify the aims of some theory or concept non-experimentally by reflecting on cases. However, insofar as philosophers are trying to give theories that connect to the ordinary language origins of philosophical topics, it seems a useful thing to know what ordinary people see (implicitly or explicitly) as the aim of relevant bits of their language.

5.3.3 So … What is Actual Causation For?

Structural causation has a clear purpose against which the quality of an inference or model may be judged: policy prediction. A structural causal model lets one say what would happen if some variable in the model were manipulated. Depending on the kind of causal structure under consideration, the result of a manipulation might be more or less specific. A causal graph will tell you what variables affect which other variables, but it will not tell you the extent or the form of the influence. A parameterized causal structure is much more informative. But either way, structural causation is informative about policy interventions.

Actual causation does not appear to be about policy prediction or control. What purpose does actual causation serve? From the armchair, it seems that actual causation is broadly diagnostic. In medicine, engineering, and quality control settings, actual causation is about fault detection. These researchers ask, "What broke or came out of alignment?" Such questions presume an ideal or normal state of affairs (in a prescriptive, not a descriptive sense) from which the actual state of affairs might deviate. In politics, economics, history, and law, actual causation is about attributing praise or blame; it is about ascribing responsibility. These researchers ask, "Who or what is responsible (either for good or for ill) for this outcome?" One might ask who or what was responsible for the Great Depression; one might ask who or what was responsible for various developments in science; or one might ask who or what was responsible for the deaths of the eleven oil rig workers in the Deepwater Horizon explosion.

The answers that one gives to these questions will surely depend on the reasons one has for asking them. Take the Deepwater explosion as an example. If one is seeking legal liability or moral culpability, then saying that the workers were killed because some concrete was defective is hardly satisfying. But knowing that the workers were killed because safety

inspectors overlooked violations of federal regulations or that the workers were killed because company executives did nothing to insure compliance with regulations answers the question satisfactorily.

5.4 CHALLENGES FOR EXPERIMENTALISTS

Investigating actual causation experimentally presents several challenges. Some of these challenges are obvious: for example, how do we identify intuitions experimentally if we cannot even be sure that we can identify them introspectively? One way to avoid the challenges of searching for intuitions is to move to something more readily observable, like linguistic behaviors.[73] Another obvious challenge for experimentation is the possible difference between snap judgments and trained, reflective judgments. If one has given up on the idea that intuitions are inflexible and infallible, then the kind of intuitive judgment that matters is probably not the initial opinion that one has when presented with a philosophical problem. Rather, what matters is the judgment one settles on after due deliberation.

A more difficult challenge for experimental investigation of actual causation, which applies equally to Socratic and Euclidean approaches, is identifying the target concept in the first place. We have already seen a kind of worry with respect to causation research: the difference between structural causation and actual causation. When one is investigating actual causation experimentally, one needs to be sure that the experiments are getting at *actual* causation, rather than *structural* causation. Lots of other distinctions are available, which are similarly troubling. For example, one might distinguish singular and generic causation, retrospective and prospective

[73] I will have more to say about the differences between fitting intuitions and fitting linguistic behaviors in Chapter 5.

causation, production and dependence, and so on. For each of these distinctions, one has to be careful to make sure that the experiments are getting at the right conceptual target.

Related to worries about whether experiments are targeting the ordinary concept of actual causation, one might worry that we just do not have a good sense of what is essential to the concept of actual causation. A number of researchers have found that moral features of causal stories make a difference to the causal judgments that people make. One reaction to such experimental results is to say that people simply fail to make the normatively correct judgments in many cases, and then the influence of moral features of a case on judgments about actual causation might be seen as nuisance influences. Paying attention to the moral features of a case might be regarded as a cognitive bias.

For example, Alicke (1992) presented subjects with a variety of similar stories. In one story, a man named John is speeding home in order to hide an anniversary present for his parents. In another story, John is driving home to hide a vial of cocaine from his parents. In both stories, he is involved in an accident on his way home. The details of the accident are changed to fill out a total of six stories. Subjects were asked to complete the sentence: The primary cause of this accident was _____. Regardless of the details of the accident, John is more likely to be judged the primary cause of the accident when he is trying to hide cocaine.

Although Alicke does not draw this conclusion, one might argue that the subjects are mistaken to pay attention to the acceptability of John's behavior in making their judgments about what caused the accident. Before doing any experimental work, I would have said that while John caused (or was a cause of) the accident in both cases, he is only responsible for the outcome in one of the two cases. After doing some experimental work, it is unclear whether I would be using the word "cause" for the same purpose as ordinary people use it, and hence, it is unclear

whether I would be right in saying that people reason counter-normatively about actual causation in such cases.

Seemingly irrelevant details sometimes matter for the judgments people make in test cases. In the examples from Alicke, the feature that makes the difference is pretty obvious, though exactly *how* it makes a difference is less clear. But in some cases, what is making the difference for ordinary judgments is not at all obvious. A further problem for experimental research on causation is an extreme sensitivity to changes in the cover story that overlays a causal structure (plus variable values). As an example of the difference that changes in cover story make, consider the following experiments.

I presented participants with three variants on a pre-emption case. In each case, there are two potential causes of an effect. (More precisely, they are preventers.) One cause runs to completion and the other does not. However, if the first had not run to completion, the second would have brought about the same effect. In all three of my cases, the second cause is a (perfect) backup for the first.

Each participant saw one of three different cases: the *Fence* case, the *Stereo* case, or the *Gravity* case. Each case consisted of a vignette followed by a prompt to select the best answer from among four choices. I presented the Fence case to 135 participants, the Stereo case to 96 participants, and the Gravity case to 49 participants who visited the Philosophical Personality website. The four vignettes, along with the answer choices, are given below:

The Fence Case

Peter and Sally like to play catch in a small park next to a highway. A high chain-link fence separates the park from the highway. One day when Peter and Sally were playing catch, Peter threw the ball a little wildly to Sally's right. Somehow Sally managed to stop the ball before it got passed her to the fence, but had Sally missed the ball, the fence would certainly have stopped it from rolling onto the highway.

Choose the best answer:

A. Sally alone prevented the ball from rolling onto the highway.
B. The fence alone prevented the ball from rolling onto the highway.
C. Both Sally and the fence prevented the ball from rolling onto the highway.
D. Neither Sally nor the fence prevented the ball from rolling onto the highway.

The Stereo Case

Peter and Sally live together in an apartment building that prohibits loud music after ten o'clock. Peter likes to listen to loud music, and he sometimes lets it play when he is not around. However, he always sets a timer to lower the volume to "two" at five till ten in order not to offend the neighbors. One day, Sally needed to study, so she turned the volume on Peter's stereo down to "two" at nine o'clock. At five till ten, the timer went off, but the volume was already set to "two" so nothing happened.

Choose the best answer:

A. Sally alone prevented the music from being loud after ten o'clock.
B. The timer alone prevented the music from being loud after ten o'clock.
C. Both Sally and the timer prevented the music from being loud after ten o'clock.
D. Neither Sally nor the timer prevented the music from being loud after ten o'clock.

The Gravity Case

Tom is six years old. One day while playing in his backyard, Tom decides to try to hit the moon with a big rubber ball he is playing with. He aims carefully and throws his ball as hard as he can. On its way up, the ball gets stuck in a tree branch. Had the ball not gotten stuck in the tree branch, gravity would certainly have stopped it from hitting the moon.

Choose the best answer:

A. The tree branch alone prevented the ball from hitting the moon.
B. Gravity alone prevented the ball from hitting the moon.
C. Both gravity and the tree branch prevented the ball from hitting the moon.
D. Neither gravity nor the tree branch prevented the ball from hitting the moon.

The raw counts for each answer by probe (along with percentages in parentheses) are given in Table 5.1.

Table 5.1: Counts for Fence, Stereo, and Gravity Cases

	A	B	C	D
Fence	66 (48.9)	10 (7.4)	49 (36.3)	10 (7.4)
Stereo	18 (18.8)	24 (25)	47 (49)	7 (7.3)
Gravity	6 (12.2)	7 (14.3)	24 (49)	12 (24.5)

The percentages from Table 5.1 are reproduced graphically in Figure 5.1.

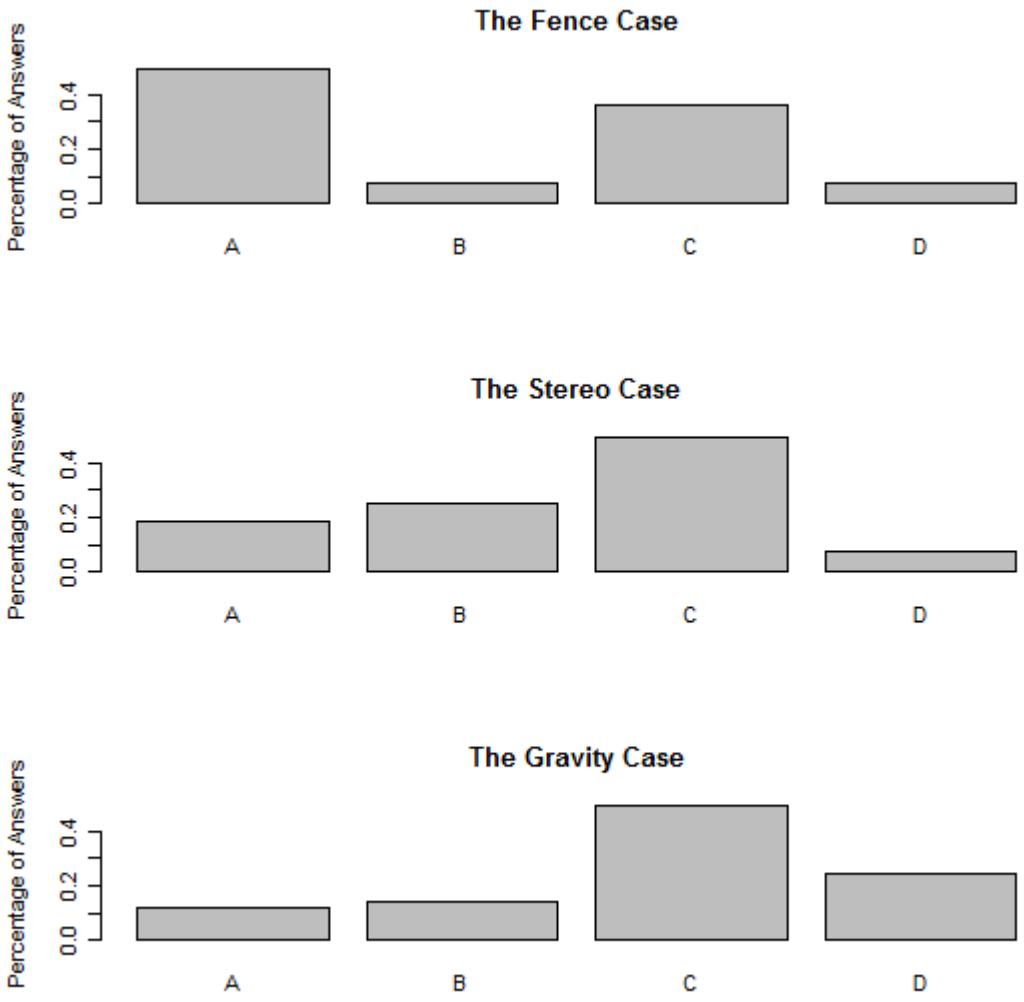

Figure 5.1: Graphical Representation of Table 5.1

Obviously, the three cases are being treated differently by participants. Or, rather, the first case is being treated differently from the second and third cases, which are being treated more or less similarly. The difference cannot consist in the perfection of the back-up. The difference cannot consist in an implicit belief that non-human, machine-like objects are better back-ups, since participants treat the fence and the stereo timer differently. What are participants picking up on that makes them treat these cases differently?

Another challenge for research about actual causation is the large variety of ways that we can talk about causation. Causal language that uses the word "cause" or related words is already highly abstract. As Anscombe (1993) notes, action words like "cut," "break," "burn," and so on have implicit causal content. But even restricting attention to explicitly causal language, the range of causal language is still very large. One might say:[74]

(5.1) $C = c$ caused $E = e$

(5.2) $C = c$ causes $E = e$

(5.3) $C = c$ is causing $E = e$

(5.4) $C = c$ is a cause of $E = e$

(5.5) $C = c$ is the cause of $E = e$

(5.6) $C = c$ is a partial cause of $E = e$

(5.7) $E = e$ was caused by $C = c$

(5.8) $E = e$ occurred because $C = c$ occurred

The list is not intended to be exhaustive.[75] Do people treat these constructions differently? In what ways? Some research has been done bearing on this problem, but not very much (Livengood and Machery 2007).

Finally, a common topic in discussion of experimental philosophy is the place of experts. Should the intuitions of experts count for more than the intuitions of non-experts, or vice versa? On the one hand, experts have extensive training, and one might think that training improves intuitions, just as training improves perception, memory, and imagination. On the other hand, experts have extensive training, and one might think that training distorts intuitions, or at least tends to make the intuitions conform to theory rather than the other way around. With respect to

[74] Recall that in cases of actual causation, the relata of the causal relation are formally described as $X = x$, where X is a random variable and x is a value of X.
[75] See Lewis 1973for a different, shorter, list.

actual causation, the dilemma seems especially pressing. One way for an expert to resist apparent counter-examples to a favored theory is to simply evaluate each new case according to the theory and give the theory's answer, come what may. Does this ability on the part of the expert make him or her a *better* source of intuitive evidence?

5.5 THE CHOICE

Researchers interested in offering a normative theory of actual causation have a choice to make. On the one hand, researchers might compare the deliverances of prospective theories to our intuitions, case by case, and accept the theory that does the best job of capturing or fitting our intuitions. Many philosophers endorse the claim that intuitions are of primary importance in philosophy. For example, Pust (2001) writes:

> Contemporary analytic philosophy is based upon intuitions. Intuitions serve as the primary evidence in the analysis of knowledge, justified belief, right and wrong action, explanation, rational action, intentionality, consciousness and a host of other properties of philosophical interest. Theories or analyses of the properties in question are attacked and defended largely on the basis of their ability to capture intuitive judgements. (227)

Other philosophers think that the use of intuitions in philosophy is an essential feature of the discipline. For example, Bealer and Strawson (1992) think that philosophy is autonomous and distinct from the sciences in virtue of its evidential use of intuitions. Even philosophers who deny that intuitions have such pride of place concede that much philosophical speculation is guided and constrained by intuitions, at least in practice (Ichikawa and Jarvis 2009; Ichikawa 2009). And so, the Socratic framework has come to be (largely) identified with philosophy as a whole.

On the other hand, researchers might try to establish with mathematical precision the instrumental value of a theory with respect to some aim or goal taken as primitive. By contrast

with the Socratic framework, the pragmatic, Euclidean approach first identifies the aim of a concept or the advantage conferred by having such a concept. Then he or she stipulates axioms designed to facilitate satisfying the aim of the concept or gaining the optimal advantage. The axioms allow the Euclidean to prove things about the degree to which the aim may be achieved, and the practical fruits of such an endeavor serve to justify the Euclidean practice.

The Socratic researcher faces the difficult challenge of establishing that intuitions (in whatever sense humans have them) track the truth about actual causation or at least that the method of reflective equilibrium succeeds in settling disagreements about actual causation. The Euclidean researcher, on the other hand, faces the difficult challenge of identifying the aim of a concept of actual causation and measuring how well or ill a given theory satisfies that aim.

CHAPTER SIX

SUMMARY AND CONCLUSIONS

I began this dissertation with two questions about the performance of U.S. students on international science and mathematics tests: What accounts for the relatively poor performance of U.S. students on international science and mathematics tests, and what should be done to improve the performance of U.S. students going forward? I noted that the two questions are fundamentally different and that cause-of-effect problems like that posed by the first question are not nearly as well-understood as problems of the second sort. I argued that actual causation is the really interesting cause-of-effect reasoning problem, and I showed that the problem of actual causation is distinct from the problems that Dawid, Wasserman, and Pearl describe (even though they all claim to be talking about the same thing). After considering Dawid's approach to cause-of-effect reasoning through concomitant variables, I recommended an approach based on structural causal models, instead. In order to make the selection of concomitant variables less arbitrary and more informative, I claimed that one should search for causal models consistent with available data and then apply a (metaphysical) definition of actual causation.

The problem with that approach is that actual causation seems to go beyond the data in an important way, and no consensus has yet appeared with respect to the correct (metaphysical) definition of actual causation, even within the structural equation tradition. I then showed that several current theories of actual causation fail to satisfy intuitions about simply voting

scenarios. I argued that the failure of the theories to satisfy these intuitions was prima facie evidence against the theories.

I then stepped back and asked how one might adjudicate between competing theories of actual causation, especially given that they seem to go beyond the data. I set out two options—one broadly rationalistic (Socratic) and one broadly pragmatistic (Euclidean). I argued that playing the Socratic game with respect to actual causation is a losing proposition. Instead of investing clever stories to elicit intuitions and then building steadily more complex theories to fit those intuitions, theorists should aim to get clear on the purpose of actual causation judgments and then build a normative theory that tells us how we *ought* to reason about actual causation.

Ordinary use of causation language is closely tied to responsibility and fault identification, at least in many cases involving agents. And so, one might think that every actual causation judgment is ultimately about assigning responsibility (or fault) for an outcome. Hitchcock and Knobe suggest that actual causation judgments depend on judgments about the default values of variables in a causal system, which are themselves determined by judgments of overall normality. And they claim that the purpose of actual causation judgments is to locate ideal places to intervene on a system.

Attempting to solve the problem of actual causation all at once via overall normality is a mistake. Instead, we should restrict attention to cases where we have a well-understood standard by which to measure the success or failure of actual causation judgments. I think that the place to look is model-based fault detection. In designed and pseudo-designed (evolved) causal systems, like mechanical and biological entities, we have a clear sense of what an actual cause is: An actual cause is whatever has to be returned to its proper function in order for the system as a whole to function properly again. The right account will still treat causation as a normative

concept. However, the kind of norm at stake is a norm of proper functioning, rather than a statistical norm or a moral norm.

APPENDIX

PROOFS OF VOTING SCENARIO RESULTS

Included in this appendix are proofs of the claims made in Chapter Three regarding the results of applying current theories of actual causation to simple voting scenarios.

A.1 TWO-CANDIDATE, SIMPLE-MAJORITY

The scenario envisioned here is very simple. Everyone must cast a vote. Each vote cast is for exactly one of two candidates, *A* and *B*. If both candidates receive the same number of votes, then the election results in a tie. Otherwise, whichever candidate receives the most votes wins the election. Thus, the election may end in a victory for one or the other of the two candidates, or it may end in a tie.

A.1.1 Hitchcock and Woodward

Consider an election in which there are $2k$ votes. (I leave it to the reader to show that elections with an odd number of votes produce identical results.) Suppose that i votes are cast for candidate *A*, and suppose without loss of generality that $2k - i < i$. Thus, candidate *A* is the actual winner of the election.

Is a vote for candidate A an actual cause of candidate A's victory? Yes. Choose a vote $V_A = A$ for candidate A.[76] There is only one path from V_A to the result of the election, and no other vote is on this path. Hence, we are free to change any of the other votes, so long as candidate A wins the election after the changes. Distribute the votes such that there are $k + 1$ votes for A (including V_A) and $k - 1$ votes for B. Now, change the value of V_A from A to B. Since such a change results in a tie (k votes for A against k votes for B), $V_A = A$ is an actual cause of A's victory. Because V_A was chosen arbitrarily, the same reasoning applies to every vote for candidate A. Hence, every vote for candidate A is an actual cause of candidate A's victory.

What about a vote for candidate B? No. No vote for candidate B is an actual cause of candidate A's victory. Choose a vote $V_B = B$ for candidate B. Again, there is only one path from V_B to the result of the election, and no other vote is on this path. Hence, we are free to change any of the other votes, so long as candidate A wins the election after the changes. However, there is no redistribution of the votes such that A is the winner of the election but would not have been the winner had V_B not been a vote for candidate B. Let r be the redistributed votes for candidate A. Since candidate A must be the winner after any redistribution, $2k - r < r$. For v_B to be an actual cause of candidate A's election, there must be k, $r \geq 0$ such that $2k - r - 1 \geq r + 1$. That is, a change in vote V_B must result in a change in the election, either to a tie or to a victory for B. But $2k - r < r \Rightarrow 2k - r < r + 2 \Rightarrow 2k - r - 1 < r + 1$. So, $V_B = B$ is not an actual cause of candidate A's election. Because V_B was chosen arbitrarily, the same reasoning applies to every vote for candidate B. Hence, no vote for candidate B is an actual cause of candidate A's victory.

[76] Given my definition of random variable, we really should code the variables so that the values are real numbers. For the sake of clarity, I have simply used the letter of the candidate, instead.

A.1.2 Halpern and Pearl

Consider an election in which there are $2k + 1$ votes. (I leave it to the reader to show that elections with an even number of votes produce identical results.) Suppose that i votes are cast for candidate A, and suppose without loss of generality that $2k + 1 - i < i$. Thus, candidate A is the actual winner of the election.

Is a vote $V_A = A$ for candidate A an actual cause of candidate A's victory? Yes. The only path from V_A to the outcome is the direct edge connecting them. Thus, we are free to set all the votes however we like so long as candidate A still wins the election if an arbitrary subset of the votes were assigned the new values while V_A retains its actual value. To satisfy (HP3), assign new values to the votes by leaving $k + 1$ votes (including V_A) at their original values A and setting k votes to B. This can always be done, since there were originally i votes for candidate A, and by supposition, $k < i$. The required redistribution of votes is shown graphically in Figure A.1 below.

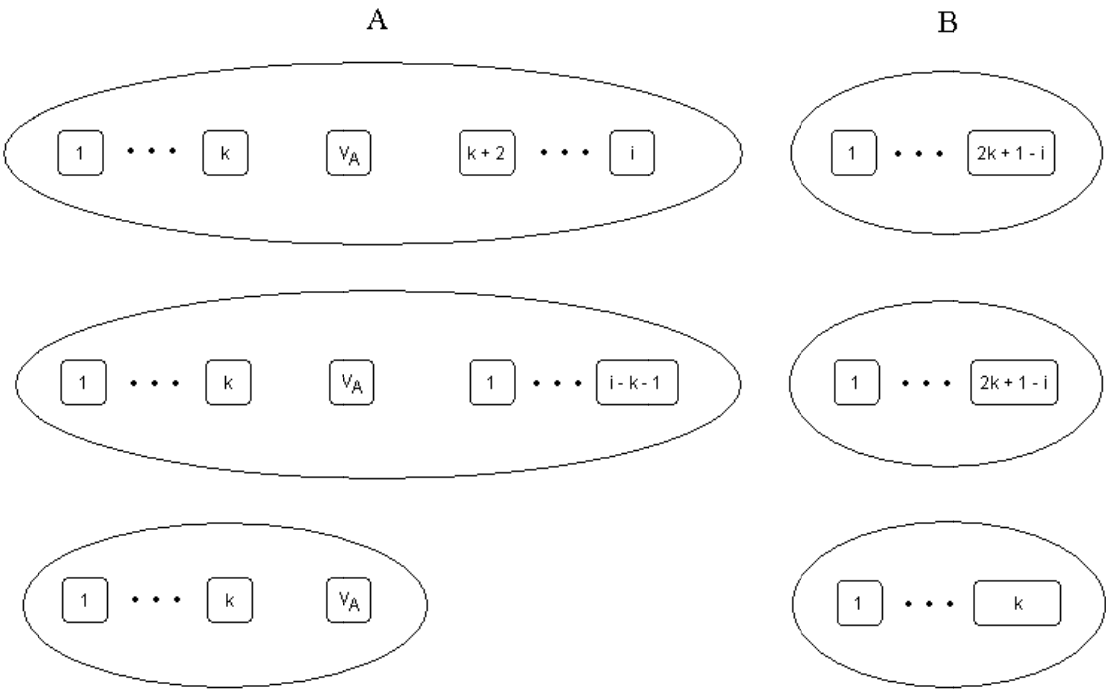

Figure A.1: First Halpern and Pearl Proof Illustration

Since setting V_A to *B* results in a victory for candidate *B*, (HP2) is satisfied as well. Hence, V_A = *A* is an actual cause of candidate *A*'s victory. Because V_A was chosen arbitrarily, the same reasoning applies to every vote for candidate *A*.

Is a vote $V_B = B$ for candidate *B* an actual cause of candidate *A*'s victory? No. No assignment of values to the votes that satisfies (HP2) can also satisfy (HP3). Let $do(\boldsymbol{V} = \boldsymbol{v}^*)$ be a proposed manipulation satisfying (HP2). Since candidate *A* won the election and the original value of V_B was *B*, the change in the outcome of the election needed in order to satisfy (HP2) must be due *solely* to changes in the values of votes for candidate *A*. Let $\boldsymbol{W}$ be an ordered tuple of the votes for *A* that were changed to votes for *B* in carrying out the manipulation to satisfy (HP2). Leaving V_B at its actual value and changing all the variables in $\boldsymbol{W}$ to their manipulated

values, candidate *B* wins the election in violation of (HP3). Hence, no vote for candidate *B* is an actual cause of candidate *A*'s victory.

A.1.3 Hall

In order to apply Hall's theory, we need to identify default values for the variables in the model. Neither a vote for candidate *A* nor a vote for candidate *B* can be thought of as the default choice. Hence, on one reading, Hall's theory can only be applied to the null reduction in forced-choice models. For null reductions of two-candidate simple-majority elections, Hall's theory reduces to simple counterfactual dependence. If the outcome of the election counterfactually depends on the actual value of some vote, then that vote is an actual cause of the outcome; otherwise, not.

However, on another reading, Hall's theory *can* be applied to the case. The issue is about how to model the scenario. One might think that even though votes must be for one of the two candidates in the actual scenario, one should *model* the scenario with three-valued variables, which include abstentions. The contrast Hall needs to draw is not between voting for candidate *A* as opposed to voting for candidate *B*; rather, the contrast he needs to draw is between voting for some candidate and not voting at all. As we will see below, if we model the scenario with three-valued variables including abstentions, then for Hall, all and only votes for candidate *A* count as actual causes of candidate *A*'s victory.

A.2 TWO-CANDIDATE, SIMPLE-MAJORITY WITH ABSTENTIONS

In this scenario, every vote cast is for exactly one of the two candidates (just like in the previous scenario). However, voters are no longer obligated to cast a vote. As before, if both candidates receive the same number of votes, then the election results in a tie. Otherwise, whichever

candidate receives the most votes wins the election. So again, the election may end in a victory for one or the other of the two candidates, or it may end in a tie.

A.2.1 Hitchcock and Woodward

Consider an election in which there are $2k$ votes. Suppose that i votes are for A, j votes are for B, and l votes are abstentions. Suppose without loss of generality that $i > j$.

Is a vote for candidate A an actual cause of candidate A's victory? Yes. Choose a vote $V_A = A$. Distribute the votes such that there is one vote for candidate A (V_A itself), zero votes for candidate B, and $2k - 1$ abstentions. Changing the value of V_A from A to B results in a victory for candidate B (zero votes for A against one vote for B), so $V_A = A$ is an actual cause of A's victory.

Is a vote for candidate B an actual cause of candidate A's victory? No. Choose a vote $V_B = B$. There is no redistribution of the votes such that A wins the election but would not after $do(V_B \neq B)$. Let r be the redistributed votes for candidate A and a be the redistributed abstentions. Since candidate A must be the winner after any redistribution, $2k - r - a < r$. For $V_B = B$ to be an actual cause of candidate A's election, there must be $a, k, r \geq 0$ such that either $2k - r - a - 2 \geq r$ (in case the vote for B is changed to a vote for A) or $2k - r - a - 1 \geq r$ (in case the vote for B is changed to an abstention). But $2k - r - a < r \Rightarrow 2k - r - a - 1 < r \Rightarrow 2k - r - a - 2 < r$. So, $V_B = B$ is not an actual cause of candidate A's election.

Are abstentions actual causes of candidate A's victory? Yes, they are. Choose an abstention, $V_{none} = 0$. Distribute the votes such that there is one vote for candidate A, zero votes for candidate B, and $2k - 1$ abstentions. Change the value of V_{none} from an abstention to a vote for B. Since such a change results in a tie (one vote for A against one vote for B), $V_{none} = 0$ is an actual cause of A's victory.

A.2.2 Halpern and Pearl

Consider an election in which there are $2k + 1$ votes. Suppose that i votes are cast for candidate A, j votes are cast for candidate B, and there are l abstentions. Suppose without loss of generality that $j < i$.

Is a vote $V_A = A$ for candidate A an actual cause of candidate A's victory? Yes. To satisfy (HP3), assign new values to the votes by leaving $j + 1$ votes (including V_A) for candidate A at their original value A, leaving all j votes for candidate B at their original value B and setting all the other votes to abstentions. The required redistribution of votes is shown graphically in Figure A.2 below.

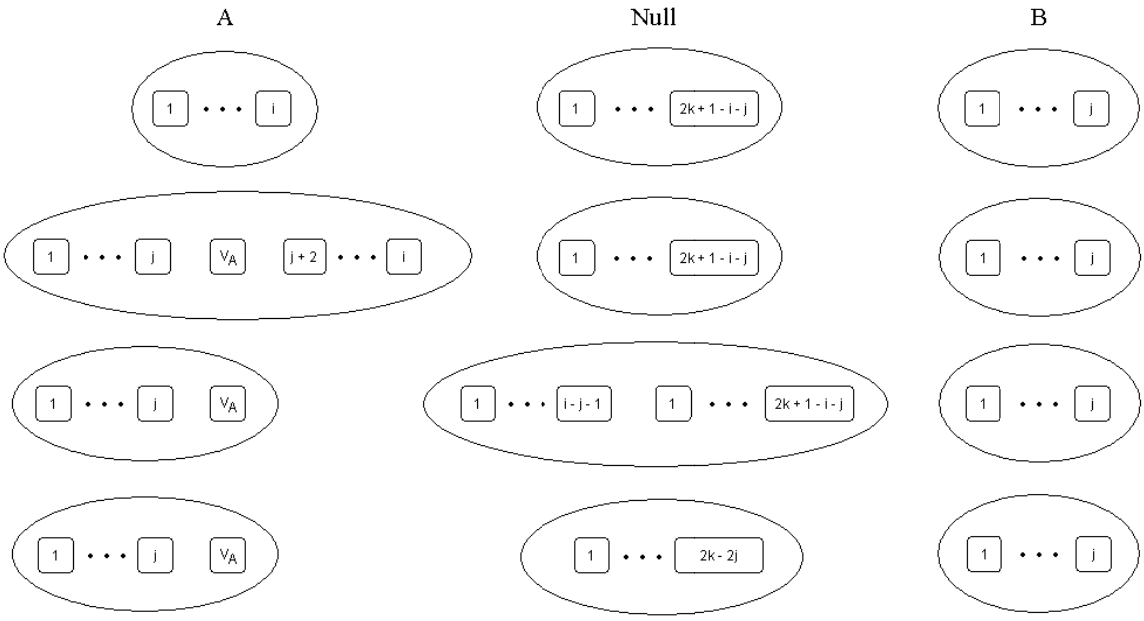

Figure A.2: Second Halpern and Pearl Proof Illustration

Since all of the votes being set to abstentions were actually either votes for A or abstentions, setting an arbitrary subset of the votes to their manipulated values leaves candidate A as the winner, so long as V_A has its actual value. Under the new assignment of values to the votes, if

the value of V_A is changed, then the election results either in a tie or a victory for candidate *B*, which satisfies (HP2). Hence, $V_A = A$ is an actual cause of *A*'s victory.

Is a vote $V_B = B$ for candidate *B* an actual cause of candidate *A*'s victory? No. No assignment of values to the votes that satisfies (HP2) can also satisfy (HP3). Let $do(\boldsymbol{V} = \boldsymbol{v}^*)$ be a proposed manipulation satisfying (HP2). Since candidate *A* won the election and the original value of V_B was *B*, the change in the outcome of the election needed in order to satisfy (HP2) must be due *solely* to changes in the values of votes for candidate *A*. Let $\boldsymbol{W}$ be an ordered tuple of the votes for *A* that were changed to votes for *B* in carrying out the manipulation to satisfy (HP2). Leaving V_B at its actual value and changing all the variables in $\boldsymbol{W}$ to their manipulated values, either candidate *B* wins the election or the election results in a tie. Either way, (HP3) is violated. Hence, no vote for candidate *B* is an actual cause of candidate *A*'s victory.

Are abstentions actual causes of candidate *A*'s victory? Yes. Choose an abstention $V_{\text{none}} = 0$. To satisfy (HP3), leave all *j* votes for candidate *B* at their actual value, leave $j + 1$ votes for candidate *A* at their actual value, and set all other votes to abstentions. Since all the votes being set to abstentions were actually either votes for candidate *A* or abstentions, setting an arbitrary subset of those votes to their manipulated values leaves candidate *A* the winner, so long as V_{none} retains its actual value. Under the new assignment, if the value of V_{none} is changed to a vote for candidate *B*, then the election results in a tie rather than a victory for candidate *A*, which satisfies (HP2). Hence, $V_{\text{none}} = 0$ is an actual cause of *A*'s victory.

A.2.3 Hall

The obvious choice for a default value of a vote in a voting scenario is abstention—at least, when that value is available. (Consequently, the default value for Elect is a tie.) All reductions of elections involve setting some votes to abstentions. Consider an election in which there are $2k$

votes. Suppose that i votes are cast for candidate A, j votes are cast for candidate B, and there are l abstentions. Further, suppose without loss of generality that $i > j$.

Is a vote for candidate A an actual cause of candidate A's victory? Yes. Choose a vote $V_A = A$ for candidate A. Set all the votes except V_A to abstentions. In this reduction, changing the value of V_A changes the result of the election. Hence, $V_A = A$ is an actual cause of Elect = A.

Is a vote for candidate B an actual cause of candidate A's victory? No. To see this, choose a vote $V_B = B$ for candidate B. In every reduction that satisfies (H2) by leaving Elect = A, the number of votes for candidate A is strictly greater than the number of votes for candidate B. Elect does not depend on V_B in any of these reductions, since the result does not change for $do(V_B = A)$ or for $do(V_B = 0)$, which are the only possible manipulations of V_B. Hence, $V_B = B$ is not an actual cause of Elect = A.

Is an abstention an actual cause of candidate A's victory? Yes. Choose a vote $V_{none} = 0$. Consider the reduction in which there is one vote for candidate A, zero votes for candidate B, and $2k - 1$ abstentions (of which V_{none} is one). In this reduction, changing V_{none} to B makes Elect = 0. Hence, $V_{none} = 0$ is an actual cause of Elect = A.

A.3 THREE-CANDIDATE, SIMPLE-PLURALITY ELECTIONS

In this scenario, every vote cast is for exactly one of the three candidates. If all three candidates receive the same number of votes or if two candidates have the same number of votes as each other and more votes than the third candidate, then the election results in a tie. Otherwise, whichever candidate receives the most votes wins the election. (In other words, a candidate need not receive the majority of the votes, and there are no run-offs.) So again, the election may end

in a victory for exactly one of the three candidates, or it may end in a tie. In order to save space, the proofs have been truncated in this section.

A.3.1 Hitchcock and Woodward

Consider an election in which there are $2k$ votes. (I leave it to the reader to show that elections with an odd number of votes produce identical results.) Suppose that i votes are for A, j votes are for B, and l votes are for C. Further, suppose without loss of generality that $i > j \geq l$.

To see that every vote for candidate A is an actual cause of candidate A's victory, we proceed as before. Choose a vote $V_A = A$. Distribute the votes such that there are $k + 1$ votes for candidate A (including V_A), $k - 1$ votes for candidate B, and no votes for candidate C. Changing the value of V_A from A to B results in a tie (k votes for A against k votes for B). Hence, $V_A = A$ is an actual cause of Elect = A.

To see that every vote for candidate B is an actual cause of candidate A's victory, choose a vote $V_B = B$. Distribute the votes such that there are k votes for A, one vote for B, and $k - 1$ votes for C. Changing the value of V_B from B to C results in a tie (k votes for A against k votes for C), so $V_B = B$ is an actual cause of Elect = A. Similarly, every vote $V_C = C$ is an actual cause of Elect = A.

Thus, according to Hitchcock's theory and according to Woodward's theory, every vote cast in an election having three (or more) candidates is an actual cause of the result of the election! Notice that this result does not in any way depend on the actual number of votes cast for each candidate. Even if no one votes for candidate C, every vote for candidate B is an actual cause of candidate A's victory.

A.3.2 Halpern and Pearl

Of the theories considered in this paper, Halpern and Pearl's theory of actual causation is the most challenging to correctly apply. For the three-candidate case, I will provide general proofs for my claims about when specific votes are causes. Suppose that i votes are for A, j votes are for B, and l votes are for C. Suppose without loss of generality that $i > j \geq l$.

Is a vote $V_A = A$ an actual cause of candidate A's victory? Yes. We need to pay attention to the difference in votes for candidate A and candidate B. We consider two cases: (1) $i - j = 2m$ for some $m \in \mathbf{N}$ and (2) $i - j = 2m + 1$ for some $m \in \mathbf{N}$. Case 1. Let $i - j = 2m$ for some $m \in \mathbf{N}$. Leave $j + m + 1$ votes for A (including V_A) at their original value, and leave all l votes for C at their original value. Set $j + m - 1$ votes to B. That is, move $m - 1$ votes from A to B. Given this distribution of votes, if we change V_A to B, the election results in a tie. Case 2. Let $i - j = 2m + 1$ for some $m \in \mathbf{N}$. Again, leave $j + m + 1$ votes for A (including V_A) at their original value, and leave all l votes for C at their original value. Set $j + m$ votes to B. That is, move m votes from A to B. Given this distribution of votes, if we change V_A to B, the election results in a win for candidate B.

Is a vote $V_B = B$ an actual cause of candidate A's victory? Yes, with the following exception: A vote $V_B = B$ is an actual cause of candidate A's victory unless candidate A wins by fewer than $j + 1$ votes, the total number of votes is odd, and there are no votes for candidate C.[77] Consider three cases: (1) $j + l > i$, (2) $j + l = i$, and (3) $j + l < i$.

Case 1. Suppose $j + l > i$. Let $m = i - l$. Leave all i votes for candidate A at their original value A and leave all l votes for candidate C at their original value C. Move $m - 1$ votes (but not

[77] In the event that $i > 2j$, $l = 0$, and $i + j = 2k + 1$ for some natural number k, the difference between i and j is odd. In order for the outcome to depend on a vote for B, enough votes have to be moved from A to C such that if V_B had been a vote for C, then C would have won or tied the election. Otherwise, there will be a subset of vote re-distributions that will result in a change in the outcome without changing the value of V_B.

including V_B) from B to C. Under the new distribution, there are i votes for A, there are $i - 1$ votes for C, and V_B is still a vote for B. If we change V_B to C, the election results in a tie.

Case 2. Suppose $j + l = i$. In this case, the total number of votes is $2i$. Leave all i votes for candidate A at their original value A and leave all l votes for candidate C at their original value C. Move all the votes for B except V_B from B to C. Under the new distribution, there are i votes for A, there are $i - 1$ votes for C, and there is one vote for B (V_B itself). If we change V_B to C, the election results in a tie.

Case 3. Suppose $j + l < i$. We need to consider two sub-cases: (a) the total number of votes is even, say $2k$, for some $k \in \mathbf{N}$ and (b) the total number of votes is odd, say $2k + 1$, for some $k \in \mathbf{N}$.

If the total number of votes is even, then $i - (j + l) = i - (2k - i) = 2i - 2k$. Let $m = 2i - 2k = 2p$ for some $p \in \mathbf{N}$. Leave all l votes for C at their original value. Move all the votes for B except V_B from B to C, and then move p votes from A to C. Under the new distribution, there are $j + l + p$ votes for A, there are $j + l + p - 1$ votes for C, and there is one vote for B (V_B itself). If we change V_B to C, the election results in a tie.

If the total number of votes is odd, then $i - (j + l) = i - (2k + 1 - i) = 2i - 2k - 1$. Let $m = 2i - 2k - 1 = 2p + 1$ for some $p \in \mathbf{N}$. Leave all l votes for C at their original value. Move $j - 2$ votes for B (not including V_B) from B to C, and then move $p + 1$ votes from A to C. As long as there was at least one vote originally cast for C, the number of votes for A is guaranteed to be at least $p + 2$ greater than the number of votes for B according to the original vote distribution, satisfying (HP3). Under the new distribution, there are $j + l + p$ votes for A, there are $j + l + p - 1$ votes for C, and there are two votes for B (including V_B). If we change V_B to C, the election results in a tie. This last case is depicted graphically in Figure A.3.

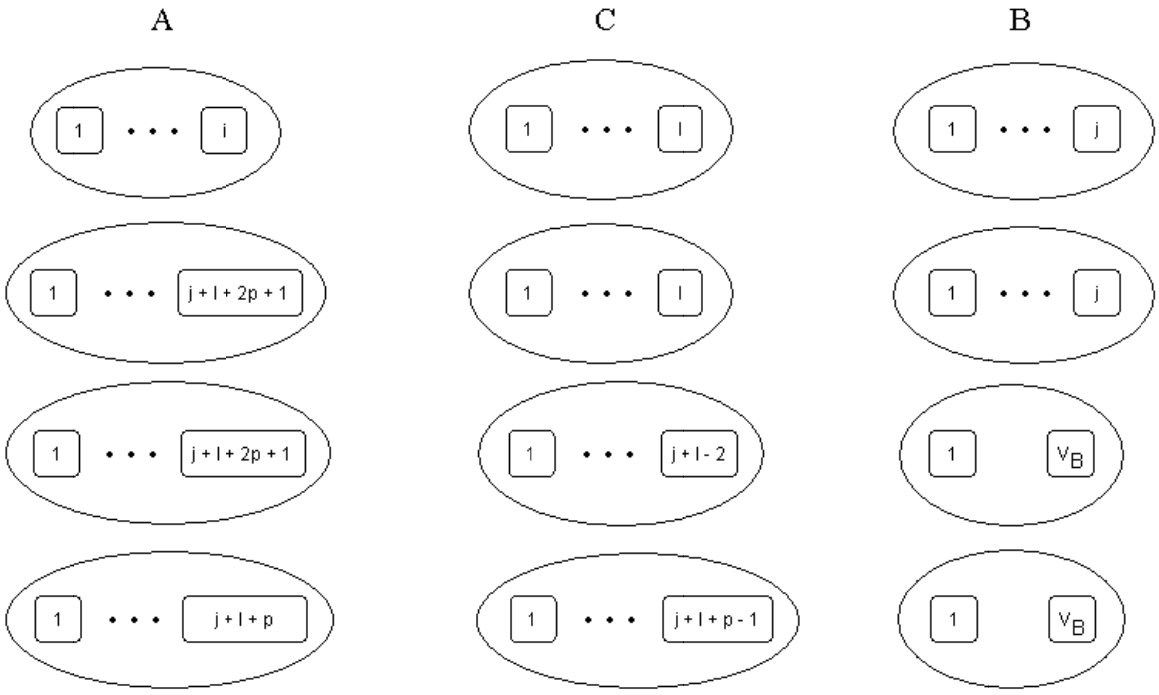

Figure A.3: Third Halpern and Pearl Proof Illustration

Is a vote $V_C = C$ an actual cause of candidate A's victory? Yes. Again, consider three cases: (1) $j + l > i$, (2) $j + l = i$, and (3) $j + l < i$.

Case 1. Suppose $j + l > i$. Let $m = i - l$. Leave all i votes for candidate A at their original value A and leave all j votes for candidate B at their original value B. Move $m - 1$ votes (but not including V_C) from C to B. Under the new distribution, there are i votes for A, there are $i - 1$ votes for B, and V_C is still a vote for C. If we change V_C to B, the election results in a tie.

Case 2. Suppose $j + l = i$. In this case, the total number of votes is $2i$. Leave all i votes for candidate A at their original value A and leave all j votes for candidate B at their original value B. Move all the votes for C except V_C from C to B. Under the new distribution, there are i votes for A, there are $i - 1$ votes for B, and there is one vote for C (V_C itself). If we change V_B to C, the election results in a tie.

Case 3. Suppose $j + l < i$. We need to consider two sub-cases: (a) the total number of votes is even, say $2k$, for some $k \in \mathbf{N}$ and (b) the total number of votes is odd, say $2k + 1$, for some $k \in \mathbf{N}$.

If the total number of votes is even, then $i - (j + l) = i - (2k - i) = 2i - 2k$. Let $m = 2i - 2k = 2p$ for some $p \in \mathbf{N}$. Leave all j votes for B at their original value. Move all the votes for C except V_C from C to B, and then move p votes from A to B. Under the new distribution, there are $j + l + p$ votes for A, there are $j + l + p - 1$ votes for B, and there is one vote for C (V_C itself). If we change V_C to B, the election results in a tie.

If the total number of votes is odd, then $i - (j + l) = i - (2k + 1 - i) = 2i - 2k - 1$. Let $m = 2i - 2k - 1 = 2p + 1$ for some $p \in \mathbf{N}$. Leave all j votes for B at their original value. Move $l - 2$ votes for C (not including V_C) from C to B, and then move $p + 1$ votes from A to B. Since $j \geq l$ by assumption, the number of votes for A is guaranteed to be at least $p + 2$ greater than the number of votes for both B and C according to the original vote distribution, satisfying (HP3). Under the new distribution, there are $j + l + p$ votes for A, there are $j + l + p - 1$ votes for B, and there are two votes for C (including V_C). If we change V_C to B, the election results in a tie.

A.3.3 Hall

Consider an election in which there are $2k$ votes. Suppose that i votes are for A, j votes are for B, and l votes are cast for C. (In order to apply Hall's theory, assume that abstention is a possible value for any vote V.) Further assume without loss of generality that $i > j \geq l$.

Is a vote for candidate A an actual cause of candidate A's victory? Yes. Choose a vote $V_A = A$. Set all the votes except V_A to abstentions. In this reduction, changing the value of V_A changes the result of the election. Hence, $V_A = A$ is an actual cause of Elect = A.

Is a vote for candidate *B* an actual cause of candidate *A*'s victory? As long as there is at least one actual vote for candidate *C*, the answer is "Yes." Choose a vote $V_B = B$. Consider the reduction in which there are two votes for *A*, one vote (namely, V_B itself) for *B*, and one vote for *C*. All other votes have been set to abstentions. In this reduction, changing V_B to a vote for *C* results in a tie. Hence, $V_B = B$ is an actual cause of Elect = *A*. The proof showing that a vote $V_C = C$ is an actual cause of Elect = *A* as long as there is at least one actual vote for *B* is a mirror image of the proof for $V_B = B$.

REFERENCES FOR THE INTRODUCTION

Aaronson, D., et al. (2007) "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25(1), 95-135.

Anderson, N. (2010) "International test score data show U.S. firmly mid-pack," *Washington Post*. http://www.washingtonpost.com/wp-dyn/content/article/2010/12/07/AR2010120701178.html

Banchero, S. (2011) "Illinois Attempts to Link Teacher Tenure to Results," *Wall Street Journal*. http://online.wsj.com/article/SB10001424052748704111504576060122295287678.html

Barron, J. (2011) "Teacher accountability bill advances in Wyoming House," *Casper Star-Tribune*. http://trib.com/news/state-and-regional/govt-and-politics/article_3cbf777a-f369-52bb-94ce-7f231cf14418.html

Berliner, D. (2006) "Our Impoverished View of Educational Reform," *Teachers College Record* 108(6), 949-995.

Berliner, D. (2009) "Poverty and Potential: Out-of-School Factors and School Success," Research Report for the Great Lakes Center for Education Research and Practice. http://bulletin.spps.org/sites/8408bc37-7c5a-435e-b738-2d321c0648bd/uploads/Report_from_Brown_Center.pdf

Boyd, D., et al. (2008) "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High-Poverty Schools," *Journal of Policy Analysis and Management* 27(4), 793-818.

Buddin, R. and G. Zamarro (2009) "Teacher qualifications and student achievement in urban elementary schools," *Journal of Urban Economics* 66, 103-115.

Burnett, K. and G. Farkas (2009) "Poverty and family structure effects on children's mathematics achievement: Estimates from random and fixed effects models," *Social Science Journal* 46, 297-318.

Calefati, J. (2011) "State Sen. Teresa Ruiz pushes new teacher tenure reform bill," *New Jersey Star-Ledger*. http://www.nj.com/news/index.ssf/2011/05/nj_democrat_pushes_new_tenure.html

Dawid, P. (2000) "Causal Inference Without Counterfactuals," *Journal of the American Statistical Association* 95, 407-424.

Etter, L. (2010) "American Teens Trail Global Peers in Math Scores," *Wall Street Journal*. http://online.wsj.com/article/SB10001424052748703471904576003842497574526.html

Gabriel, T. and S. Dillon (2011) "G.O.P. Governors Take Aim at Teacher Tenure," *New York Times*. http://www.nytimes.com/2011/02/01/us/01tenure.html

Ghianni, T. (2011) "Tennessee legislature passes bill changing teacher tenure rules," *Reuters*. http://www.reuters.com/article/2011/03/24/us-tennessee-teachers-idUSTRE72N7DJ20110324

Hitchcock, C. and J. Knobe (forthcoming) "Cause and Norm," *Journal of Philosophy*.

Kane, T. and D. Staiger (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," *National Bureau of Economic Research*, Working Paper 14607. http://www.nber.org/papers/w14607

Krashen, S. (2004) *The Power of Reading*. Portsmouth: Heinemann and Westport.

Krashen, S. (2010) "How poverty affected U.S. PISA scores," *Washington Post*. http://voices.washingtonpost.com/answer-sheet/research/how-poverty-affected-us-pisa-s.html

Krashen, S., et al. (2010) "An Analysis of the PIRLS (2006) Data: Can the School Library Reduce the Effect of Poverty on Reading Achievement?" *CSLA (California School Library Association) Journal* 34(1), 26-28.

Mahler, J. (2011) "The Deadlocked Debate Over Education Reform," *New York Times*. http://www.nytimes.com/2011/04/10/weekinreview/10reform.html?_r=3&hp

Mazzei, P. (2011) "Florida House approves teacher tenure law," *Miami Herald*. http://www.miamiherald.com/2011/03/16/2118711/teacher-reform-bill-gets-final.html

McCabe, C. (2010) "The Economics Behind International Education Rankings," *NEAToday*. http://neatoday.org/2010/12/09/a-look-at-the-economic-numbers-on-international-education-rankings/

Myers, S., et al. (2004) "The Effect of School Poverty on Racial Gaps in Test Scores: The Case of the Minnesota Basic Standards Tests," *Journal of Negro Education* 73(1), 81-98.

Neuman, S. and D. Celano (2001) "Access to print in low-income and middle-income communities: An ecological study of four neighborhoods," *Reading Research Quarterly* 36(1), 8-26.

OECD (2010) PISA *2009 Results: Executive Summary*. http://www.oecd.org/dataoecd/34/60/46619703.pdf

Payne, K. and B. Biddle (1999) "Poor School Funding, Child Poverty, and Mathematics Achievement," *Educational Researcher* 28(6), 4-13.

Payne, K. and B. Biddle (2000) "Funding, Poverty, and Mathematics Achievement: A Rejoinder to Sarah E. Turner," *Educational Researcher* 29(7), 27-29.

Peterson, P. (2010) "Charter Schools and Student Performance," *Wall Street Journal*. http://online.wsj.com/article/SB10001424052748703909804575123470465841424.html

Ravitch, D. (2011) "The Myth of Charter Schools," *New York Review of Books*. http://www.nybooks.com/articles/archives/2010/nov/11/myth-charter-schools/?pagination=false

Resmovits, J. (2011a) "Teacher Tenure Under Fire From Statehouses," *Huffington Post*. http://www.huffingtonpost.com/2011/05/12/teacher-tenure-under-fire-state-legislatures_n_861279.html

Resmovits, J. (2011b) "Alabama House Passes Bill That Maintains Teacher Tenure But Dilutes Its Protections," *Huffington Post*. http://www.huffingtonpost.com/2011/05/26/alabama-house-passes-teacher-tenure-bill_n_867585.html

Riddile, M. (2010) "PISA: It's Poverty Not Stupid," *The Principal Difference*. http://nasspblogs.org/principaldifference/2010/12/pisa_its_poverty_not_stupid_1.html

Sirin, S. (2005) "Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research," *Review of Educational Research* 75(3), 417-453.

Turner, S. (2000) "A Comment on 'Poor School Funding, Child Poverty, and Mathematics Achievement,'" *Educational Researcher* 29(5), 15-18.

Will, G. (2011) "US Schools Get Failing Grade," *Newsmax*. http://www.newsmax.com/GeorgeWill/arne-duncan-education/2011/01/28/id/384219

REFERENCES FOR CHAPTER ONE

Dawid, P. (2000) "Causal Inference Without Counterfactuals," *Journal of the American Statistical Association* 95, 407-424.

Dawid, P. (2000) "Causal Inference Without Counterfactuals: Rejoinder," *Journal of the American Statistical Association* 95, 444-448.

Dawid, P. (2006) "Counterfactuals, Hypotheticals and Potential Responses: A Philosophical Examination of Statistical Causality," *Research Report No. 269*, Department of Statistical Science, University College London, June 20, 2006.

Dawid, P. (2007) "Fundamentals of Statistical Causality," *Research Report No. 279*, Department of Statistical Science, University College London, September 17, 2007.

Gelman, A. (2010) "Causality and Statistical Learning," *American Journal of Sociology*, pre-print at http://arxiv.org/ftp/arxiv/papers/1003/1003.2619.pdf

Goodman, N. (1983) *Fact, Fiction and Forecast*, 4th Edition. Cambridge: Harvard University Press.

Hall, N. (2007) "Structural Equations and Causation," *Philosophical Studies* 132, 109-136.

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

Hitchcock, C. (1995) "The Mishap at Reichenbach Fall: Singular vs. General Causation," *Philosophical Studies* 78, 257-291.

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

Holland, P. (1986) "Statistics and Causal Inference," *Journal of the American Statistical Association* 81, 945-960.

Holland, P. (1993) "Which Comes First, Cause or Effect?" Chapter 9 in *A Handbook for Data Analysis in the Behavioral Sciences*. Edited by G. Keren and C. Lewis. Hillsdale: Lawrence Erlbaum.

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

Lewis, D. (1973b) "Causation," *Journal of Philosophy* 70, 556-567.

Lewis, D. (1979) "Counterfactual Dependence and Time's Arrow," *Noûs* 13, 455-476.

Neyman, J. (1990 [1923]) "On the applicability of probability theory to agricultural experiments," *Statistical Science* 5(4), 465-480.

Pearl, J. (2000) "Comment on Causal Inference Without Counterfactuals," *Journal of the American Statistical Association* 95, 428-431.

Rubin, D. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66(5), 688-701.

Rubin, D. (1976) "Inference and missing data," *Biometrika* 63, 581-592.

Rubin, D. (1977) "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics* 2, 1-26.

Rubin, D. (1978) "Bayesian inference for causal effects: the role of randomization," *Annals of Statistics* 7, 34-58.

Rubin, D. (1997) "Estimating Causal Effects from Large Data Sets Using Propensity Scores," *Annals of Internal Medicine*127(8S), 757-763.

Rubin, D. (2006) *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

Schaffer, J. (2003) "Overdetermining Causes," *Philosophical Studies* 114, 23-45.

Shafer, G. (2000) "Comment on Causal Inference Without Counterfactuals," *Journal of the American Statistical Association* 95, 438-442.

Spirtes, P. et al. (2000) *Causation, Prediction, and Search*, Second Edition. Cambridge: MIT Press.

Wasserman, L. (2000) "Comment on Causal Inference Without Counterfactuals," *Journal of the American Statistical Association* 95, 442-443.

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

REFERENCES FOR CHAPTER TWO

Bollen, K. (1989) *Structural Equations with Latent Variables*. John Wiley & Sons.

Eberhardt, F. and R. Scheines (2007) "Interventions and Causal Inference," *Philosophy of Science* 72, 981-995.

Eells, E. and E. Sober (1983) "Probabilistic Causality and the Question of Transitivity," *Philosophy of Science* 50(1), 35-57.

Freedman, D. (2005) *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

Freedman, D. (2010) *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Edited by D. Collier, et al. Cambridge: Cambridge University Press.

Frey, B. (2003) "Extending Factor Graphs so as to Unify Directed and Undirected Graphical Models," *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*.

Glymour, C. et al. (2010) "Actual Causation: A Stone Soup Essay," *Synthese* 175, 169-192.

Glymour, C. and F. Wimberly (2007) "Actual Causes and Thought Experiments," in *Causation and Explanation*, 43-67. Edited by Campbell, O'Rourke, and Silverstein. Cambridge: MIT Press.

Hall, N. (2007) "Structural Equations and Causation," *Philosophical Studies* 132, 109-136.

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

Halpern, J. (2008) "Defaults and Normality in Causal Structures,"

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

Hitchcock, C. (2001b) "Causal Generalizations and Good Advice," *Monist* 84(2), 222-246.

Hitchcock, C. (2007a) "Prevention, Preemption, and the Principle of Sufficient Reason," *The Philosophical Review* 116(4), 495-532.

Hitchcock, C. (2007b) “Three Concepts of Causation,” *Philosophy Compass* 2/3, 508-516.

Hitchcock, C. (2009) “Structural equations and causation: six counterexamples,” *Philosophical Studies* 144, 391-401.

Hitchcock, C. and J. Knobe (forthcoming) “Cause and Norm,” *The Journal of Philosophy*.

Holland, P. (1986) “Statistics and Causal Inference,” *Journal of the American Statistical Association* 81, 945-960.

Kline, R. (1998) *Principles and Practice of Structural Equation Modeling*. New York: Guilford.

Pearl, J. (2000) *Causality*: *Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearl, J. and E. Bareinboim (2010) “Transportability across studies: A formal approach,” *Technical Report R-372*.

Simons, L. et al. (2009) “The Effect of Religion on Risky Sexual Behavior among College Students,” *Deviant Behavior* 30(5), 467-485.

Spirtes, P. (1995) “Directed Cyclic Graphical Representations of Feedback Models,” *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. by Philippe Besnard and Steve Hanks, Morgan Kaufmann Publishers, Inc., San Mateo.

Spirtes, P. et al. (2000) *Causation, Prediction, and Search*. Cambridge: MIT Press.

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

REFERENCES FOR CHAPTER THREE

Baumgartner, M. (2008) "Regularity Theories Reassessed," *Philosophia* 36, 327-354.

Bollen, K. (1989) *Structural Equations with Latent Variables*. John Wiley & Sons.

Casella, G. and R. Berger (2002) *Statistical Inference*, Second Edition. Duxbury.

Collins, J. et al. (2004) *Causation and Counterfactuals*. Cambridge: MIT Press.

Dowe, P. (2000) *Physical Causation.* New York: Cambridge University Press, 2000.

Freedman, D. (2005) *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

Glymour, C. et al. (2010) "Actual Causation: A Stone Soup Essay," *Synthese* 175, 169-192.

Glymour, C. and F. Wimberly (2007) "Actual Causes and Thought Experiments," in *Causation and Explanation*, 43-67. Edited by Campbell, O'Rourke, and Silverstein. Cambridge: MIT Press.

Goldman, A. (1999) "Why Citizens Should Vote: A Causal Responsibility Approach," *Social Philosophy and Policy* 16(2), 201-217.

Hall, N. (2007) "Structural Equations and Causation," *Philosophical Studies* 132, 109-136.

Hall, N. (2004) "Two Concepts of Causation," in *Causation and Counterfactuals*, eds. Collins, Hall, and Paul, Cambridge: MIT Press.

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

Hart, H. and T. Honore (2002) *Causation in the Law*, 2nd Edition. Oxford: Clarendon Press.

Hitchcock, C. and J. Knobe (forthcoming) "Cause and Norm," *The Journal of Philosophy*.

Hitchcock, C. (2009) "Structural equations and causation: six counterexamples," *Philosophical Studies* 144, 391-401.

Hitchcock, C. (2007) “Prevention, Preemption, and the Principle of Sufficient Reason,” *The Philosophical Review* 116(4), 495-532.

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

Hitchcock, C. (1995) “The Mishap at Reichenbach Fall: Singular vs. General Causation,” *Philosophical Studies* 78, 257-291.

Holland, P. (1986) “Statistics and Causal Inference,” *Journal of the American Statistical Association* 81, 945-960.

Kallenberg, O. (2002) *Foundations of Modern Probability*, Second Edition. New York: Springer-Verlag.

Kline, R. (1998) *Principles and Practice of Structural Equation Modeling*. New York: Guilford.

Livengood, J. and E. Machery (2007) “The Folk Probably Don’t Think What You Think They Think: Experiments on Causation by Absence,” *Midwest Studies in Philosophy* 31, 107-127.

Mackie, J. (1965) “Causes and Conditions,” *American Philosophical Quarterly* 2(4), 245-264.

Megill, A. (2007) *Historical Knowledge, Historical Error*: *A Contemporary Guide to Practice*. Chicago: University of Chicago Press.

Menzies, P. and H. Price (1993) “Causation as a Secondary Quality,” *British Journal for the Philosophy of Science* 44, 187–203.

Norton, J. (2003) “Causation as Folk Science,” *Philosophers’ Imprint* 3(4)

Pearl, J. (2000) *Causality*: *Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Salmon, W. (1998) *Causality and Explanation.* New York: Oxford University Press.

Sartorio, C. (2004) “How to be Responsible for Something without Causing It,” *Philosophical Perspectives* 18, 315-336.

Schaffer, J. (2003) “Overdetermining Causes,” *Philosophical Studies* 114, 23-45.

Tomer, A. (2003) “A short history of structural equation models,” *Structural Equation Modeling*. B. Pugesek, editor. West Nyack: Cambridge University Press.

Wolff, P. (2007) “Representing Causation,” *Journal of Experimental Psychology: General* 136(1), 82-111.

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

REFERENCES FOR CHAPTER FOUR

Alicke, Mark (1992). "Culpable Causation." *Journal of Personality and Social Psychology*, 36: 368–378.

Beebee, H. (2004) "Causing and Nothingness," in *Causation and Counterfactuals*, edited by Collins et al., 291-308.

Bollen, K. (1989) *Structural Equations with Latent Variables*. New York: Wiley.

Collins, J. et al., eds. (2004) *Causation and Counterfactuals*. Cambridge: MIT Press.

Glymour, C. and F. Wimberly (2007) "Actual Causes and Thought Experiments," in *Causation and Explanation*, 43-67. Edited by Campbell, O'Rourke, and Silverstein. Cambridge: MIT Press.

Hacker, P. (2009) "Critical Studies: A Philosopher of Philosophy," *Philosophical Quarterly* 59(235), 337-348.

Hall, N. (2007) "Structural Equations and Causation," *Philosophical Studies* 132, 109-136.

Hall, N. (2004) "Rescued from the Rubbish Bin," *Philosophy of Science* 71, 1107-1114.

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

Hitchcock, Christopher and Joshua Knobe (forthcoming). "Cause and Norm." *Journal of Philosophy*.

Hitchcock, C. (2009) "Structural equations and causation: six counterexamples," *Philosophical Studies* 144, 391-401.

Hitchcock, C. (2007a) "Prevention, Preemptions, and the Principle of Sufficient Reason," *Philosophical Review* 116(4), 495-531.

Hitchcock, C. (2007b) "Three Concepts of Causation," *Philosophy Compass* 2/3, 508-516.

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

Holland, P. (1986) "Statistics and Causal Inference," *Journal of the American Statistical Association* 81, 945-960.

Ichikawa, J. (2009) "Explaining Away Intuitions," *Studia Philosophica Estonica* 2(2), 94-116.

Kline, R. (1998) *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press.

Knobe, Joshua (2006). *Folk Psychology, Folk Morality*. Dissertation.

Knobe, Joshua and Ben Fraser (2008). "Causal judgments and moral judgment: Two experiments." In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447. Cambridge: MIT Press.

Lewis, D. (2004) "Causation as Influence," in *Causation and Counterfactuals*, edited by Collins et al., 75-106.

Lewis, D. (1986) *Philosophical Papers*, Volume II. Oxford: Oxford University Press.

Menzies, P. (1996) "Probabilistic Causation and the Pre-emption Problem," *Mind* New Series 105(417), 85-117.

Pearl, J. (2000) *Causality*: *Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Roxborough, Craig and Jill Cumby (2009). "Folk psychological concepts: Causation." *Philosophical Psychology*, 22(2): 205–213.

Sytsma, J., J. Livengood, and D. Rose (ms). "Two Types of Typicality."

Weatherson, B. (2003) "What Good are Counterexamples?" *Philosophical Studies* 115, 1-31.

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

REFERENCES FOR CHAPTER FIVE

Anscombe, G. (1993) "Causality and Determination," in *Causation*, edited by Sosa and Tooley.

Baumgartner, M. (2008) "Regularity Theories Reassessed," *Philosophia* 36(3), 327-354.

Bealer, G. and P. Strawson (1992) "The Incoherence of Empiricism," *Proceedings of the Aristotelian Society* 66, 99-143.

Bealer, G. (1998) "Intuition and the Autonomy of Philosophy," in *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, edited by DePaul and Ramsey.

Bealer, G. (2002) "Modal Epistemology and the Rationalist Renaissance," in *Conceivability and Possibility*, edited by Gendler and Hawthorne.

Beebee, H. (2006) "Does Anything Hold the Universe Together?" *Synthese* 149, 509-533.

Beebee, H. (2007) "Hume on Causation," in *Causation, Physics, and the Constitution of Reality*: *Russell's Republic Revisited*, edited by Price and Corry.

Beebee, H., et al., eds. (2009) *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Cartwright, N. (1979) "Causal Laws and Effective Strategies," *Noûs* 13(4), 419-437.

Cartwright, N. (1996) "Fundamentalism vs. the Patchwork of Laws," in *The Philosophy of Science*, edited by Papineau.

Craver, C. (2007) *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

DePaul, M. and W. Ramsey, eds. (1998) *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Lanham: Rowman & Littlefield.

Dorr, C. (2010) "Review of *Everything Must Go: Metaphysics Naturalized*," *Notre Dame Philosophical Reviews*. http://ndpr.nd.edu/review.cfm?id=19947

Dowe, P. (1992) "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory," *Philosophy of Science* 59, 195-216.

Dowe, P. (2000) *Physical Causation*. Cambridge: Cambridge University Press.

Dowe, P. (2009) "Causal Process Theories," in *The Oxford Handbook of Causation*, edited by Beebee, et al.

Dummett, M. and A. Flew (1954) "Symposium: Can an Effect Precede its Cause?" *Proceedings of the Aristotelian Society* 28, 27-62.

Eells, E. (1991) *Probabilistic Causality*. Cambridge: Cambridge University Press.

Evans, J. (2003) "In two minds: dual-process accounts of reasoning," *TRENDS in Cognitive Science* 17(10), 454-459.

Gendler, T. and J. Hawthorne, eds. (2002) *Conceivability and Possibility*. Oxford: Clarendon Press.

Gilovich, T., et al. (2002) *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Glennan, S. (1996) "Mechanisms and the Nature of Causation," *Erkenntnis* 44(1), 49-71.

Glennan, S. (2009) "Mechanisms," in *The Oxford Handbook of Causation*, edited by Beebee, et al.

Gopnik, A. (1993) "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality," *Behavioral and Brain Sciences* 16, 1-14.

Hausman, D. (1986) "Causation and Experimentation," *American Philosophical Quarterly* 23(2), 143-154.

Hitchcock, C. (1993) "A Generalized Probabilistic Theory of Causal Relevance," *Synthese* 97(3), 335-364.

Hitchcock, C. (2007) "How to Be a Causal Pluralist," in *Thinking About Causes: From Greek Philosophy to Modern Physics*, edited by Machamer and Wolters.

Ichikawa, J. (2009) "Explaining Away Intuitions," *Studia Philosophica Estonica* 2.2, 94-116.

Ichikawa, J. and B. Jarvis (2009) "Thought-experiment intuitions and truth in fiction," *Philosophical Studies* 142, 221-246.

Kahneman, D. and S. Frederick (2002) "Representativeness revisited: Attribute substitution in intuitive judgment," in *Heuristics and biases: The psychology of intuitive judgment*, edited by Gilovich, et al.

Kahneman, D. (2003) "A Perspective on Judgment and Choice: Mapping Bounded Rationality," *American Psychologist* 58(9), 697-720.

Lewis, D. (1973) "Causation," *Journal of Philosophy* 70, 556-567.

Lewis, D. (1986) *Philosophical Papers*. Oxford: Oxford University Press.

Lewis, D. (2000) "Causation as Influence," *Journal of Philosophy* 97, 182-197.

Ludwig, K. (2007) "The Epistemology of Thought Experiments: First Person versus Third Person Approaches," *Midwest Studies in Philosophy* 31, 128-159.

Machamer, P., et al. (2000) "Thinking About Mechanisms," *Philosophy of Science* 67(1), 1-25.

Machamer, P. and G. Wolters, eds. (2007) *Thinking About Causes: From Greek Philosophy to Modern Physics*. Pittsburgh: University of Pittsburgh Press.

Machery, E. et al. (2004) "Semantics, Cross-Cultural Style," *Cognition* 92, B1-B12.

Mackie, J. (1965) "Causes and Conditions," *American Philosophical Quarterly* 2(4), 245-264.

Mackie, J. (1974) *The Cement of the Universe*. Oxford: Clarendon Press.

Menzies, P. and H. Price (1993) "Causation and a Secondary Quality," *British Journal for the Philosophy of Science* 44, 187-203.

Paul, L. (2009) "Counterfactual Theories," in *The Oxford Handbook of Causation*, edited by Beebee, et al.

Peirce, C. (1992) *The Essential Peirce: Selected Philosophical Writings*, Volume 1 (1867-1893). Edited by Houser and Kloesel. Bloomington: Indiana University Press.

Price, H. (1991) "Agency and Probabilistic Causality," *British Journal for the Philosophy of Science* 42, 157-176.

Price, H. (2007) "Causal Perspectivalism," in *Causation, Physics, and the Constitution of Reality*: *Russell's Republic Revisited*, edited by Price and Corry.

Price, H. and R. Corry, eds. (2007) *Causation, Physics, and the Constitution of Reality*: *Russell's Republic Revisited*. Oxford: Clarendon Press.

Psillos, S. (2009) "Regularity Theories," in *The Oxford Handbook of Causation*, edited by Beebee, et al.

Psillos, S. (2010) "Causal Pluralism," in *Worldviews, Science and Us: Studies of Analytical Metaphysics*, edited by Vanderbeeken and D'Hooghe.

Pust, J. (2001) “Against Explanationist Skepticism regarding Philosophical Intuitions,” *Philosophical Studies* 106(3), 227-258.

Rawls, J. (1971) *A Theory of Justice*. Cambridge: Harvard University Press.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Salmon, W. (1994) “Causality without Counterfactuals,” *Philosophy of Science* 61, 297-312.

Salmon, W. (1997) “Causality and Explanation: A Reply to Two Critiques,” *Philosophy of Science* 64, 461-477.

Schwitzgebel, E. (2004) “Introspective training apprehensively defended: Reflections on Titchener's lab manual,” *Journal of Consciousness Studies* 11(7-8), 58-76.

Schwitzgebel, E. (2007) “No unchallengeable epistemic authority, of any sort, regarding our own conscious experience—contra Dennett?” *Phenomenology and the Cognitive Sciences* 6, 107–113.

Schwitzgebel, E. (2008) “The Unreliability of Naïve Introspection,” *Philosophical Review* 117, 245-273.

Shoemaker, S. (1988) “On Knowing One’s Own Mind,” *Philosophical Perspectives* 2, 183-209.

Sloman, S. (2002) “Two Systems of Reasoning,” in *Heuristics and biases: The psychology of intuitive judgment*, edited by Gilovich et al.

Sosa, E. and M. Tooley, eds. (1993) *Causation*. Oxford: Oxford University Press.

Sosa, E. (1998) “Minimal Intuition,” in *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, edited by DePaul and Ramsey.

Stich, S. (1998) “Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity,” in *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, edited by DePaul and Ramsey.

Suppes, P. (1970) *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company.

Sytsma, J. and J. Livengood (forthcoming) “A New Perspective Concerning Experiments on Semantic Intuitions,” *Australasian Journal of Philosophy*.

Vanderbeeken, R. and B. D’Hooghe, eds. (2010) *Worldviews, Science and Us: Studies of Analytical Metaphysics*. Singapore: World Scientific Publishing.

Weatherson, B. (2003) "What Good Are Counterexamples?" *Philosophical Studies* 115, 1-31.

Weinberg, J. (2007) "How to Challenge Intuitions Empirically Without Risking Skepticism," *Midwest Studies in Philosophy* 31, 318-343.

Williamson, J. (2009) "Probabilistic Theories," in *The Oxford Handbook of Causation*, edited by Beebee, et al.

Williamson, T. (2004) "Philosophical 'Intuitions' and Scepticism about Judgement," *Dialectica* 58, 109-153.

Williamson, T. (2007) *The Philosophy of Philosophy*. Oxford: Blackwell.

Williamson, T. (2009) "Précis of *The Philosophy of Philosophy*," *Philosophical Studies* 145, 431-434.

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

Woodward, J. (2009) "Agency and Interventionist Theories," in *The Oxford Handbook of Causation*, edited by Beebee, et al.