A Structural Model Interpretation of Wright's NESS Test

A Thesis Submitted to the College of

Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the Degree of Master of Science

in the Department of Computer Science

University of Saskatchewan

Saskatoon

By
Richard Baldwin

Keywords: structural equation models, actual causation, NESS test

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5A9

A Structural Model Interpretation of Wright's NESS Test

(M.Sc. Thesis)

Candidate: Richard Baldwin

Supervisor: Eric Neufeld

Summer 2003

ABSTRACT

Although understanding causation is an essential part of nearly every problem domain, it has resisted formal treatment in the languages of logic, probability, and even statistics. Autonomous artificially intelligent agents need to be able to reason about cause and effect. One approach is to provide the agent with formal, computational notions of causality that enable the agent to deduce cause and effect relationships from observations. During the 1990s, formal notions of causality were pursued within the AI community by many researchers, notably by Judea Pearl. Pearl developed the formal language of structural models for reasoning about causation. Among the problems he addressed in this formalism was a problem common to both AI and law, the attribution of causal responsibility or actual causation. Pearl and then Halpern and Pearl developed formal definitions of actual causation in the language of structural models.

Within the law, the traditional test for attributing causal responsibility is the counterfactual "but-for" test, which asks whether, but for the defendant's wrongful act, the injury complained of would have occurred. This definition conforms to common intuitions regarding causation in most cases, but gives non-intuitive results in more complex situations where two or more potential causes are present. To handle such situations, Richard Wright defined the NESS Test. Pearl claims that the structural language is an appropriate language to capture the intuitions that motivate the NESS test. While Pearl's structural language is adequate to formalize the NESS test, a recent result of Hopkins and Pearl shows that the Halpern and Pearl definition fails to do so, and this thesis develops an alternative structural definition to formalize the NESS test.

ACKNOWLEDGEMENTS

I would like to acknowledge the following individuals:

- My thesis advisor, Prof. Eric Neufeld, who with generous patience somehow managed to get a thesis out of an always self-doubting, forever procrastinating, anything but ideal student. Water from stone; medals have been awarded for less.

- Prof. Gord McCalla, currently a member of my thesis committee, was my original thesis advisor. The many hours I spent with Prof. McCalla discussing (or, trying to avoid discussing) my thesis work were among the most rewarding of my time in Saskatoon. I will always regret not being able to bring that thesis work to fruition and always be grateful for his support.

- Prof. Michael Horsch, the third member of my thesis committee, provided comments and suggestions with respect to both my thesis proposal and final thesis that were very helpful in preparation of the final document.

- Prof. Julita Vassileva generously stepped in to act as advisor for a required research course when Prof. McCalla was away on sabbatical among the incomprehensible antipodeans.

- Jan Thompson, the Department of Computer Science Graduate Correspondent, was always available to do what was administratively necessary to (somehow) keep me going as student, and cheerfully too.

- The family Epsilon, my friends Raja, Bobby, and their son Ricky Spandan. Raja got me into this academic adventure while Spandan tried hard to get me to finish with his continuing encouragement to do my “homework” so I could come and play with him.

- Finally, my parents for being my parents, not an especially easy task. The fact that they are there for me whenever I need them is just a bonus.

TABLE OF CONTENTS

LIST OF FIGURES

1. Introduction

As Ashley (1990, p. 2) writes,

> The legal domain [is] an interesting one from the viewpoint of AI research. The legal domain is midway between logical or mathematical domains that are amenable to computer science techniques and the domains of commonsense reasoning and ordinary discourse that AI so wishes to tackle. Studying how knowledge is structured in this intermediately formalized field may lead to useful insights.

Causation is an important and difficult concept for both law and AI. For example, formalizing commonsense reasoning has been a part of the programme of AI from its beginning (see McCarthy, 1990) and causal knowledge plays a central role in commonsense reasoning about action and change. Causation is important for intelligent, autonomous agents who reason about the effects of their actions or generate explanations for events occurring in their environment (see e.g. Pearl, 1994; Halpern and Pearl, 2000). Ortiz (1999) characterizes problems facing agents whose solution involves causality as ones of prediction, planning, diagnosis, induction, and what he calls *the problem of causal attribution*; that is, given an agent's representation of the state of the (its) world, the agent's causal (law-like) and non-causal (definitional and constraint) knowledge about the world, and the description of separate events, determining what, if any, causal connection exists between those events. This is just the problem faced by a group of human agents, a jury, in deciding whether liability (guilt) attaches to a defendant in a legal case.

1.1 Actual Causation in Law—The "But-For" Test

In law the issue of causation is complex and controversial (see e.g. Hart and Honore, 1985; Honore, 2001). A significant aspect of the issue concerns the nature of the relationship that must be established between a defendant's (legally) wrongful act or failure to act and the harm complained of by the plaintiff (in a civil case) or the Crown

(in a criminal case). According to traditional theories of corrective justice, for a defendant to be held liable for his wrongful conduct that conduct must have contributed to the injury complained of, that is, it must have been a *cause-in-fact* or *actual cause* of the harm.

The generally accepted test for determining actual causation is a counterfactual necessity test, the "but-for" test. If the specified injury would not have occurred but-for the defendant's wrongful conduct then actual causation is established. The but-for test assumes that it is somehow possible to remove the defendant's wrongful conduct from the scenario describing the occurrence of the injury and determining whether the injury would have still occurred. As far as the courts are concerned, the performance of this test requires no special skills or knowledge and is therefore assigned to the trier of fact (the jury in a jury trial, otherwise the presiding judge). However, the test is not comprehensive. It fails in circumstances where the scenario describing the injury includes other potential causes that would have brought about the specified injury in the absence of the defendant's wrongful conduct. These are known as cases of *overdetermined* causation.

Wright (1975, pp. 1775-76) partitions cases of overdetermined causation into *preemptive* causation and *duplicative* causation cases. In preemptive causation cases, the effect of other potential causes is preempted by the effect of the defendant's wrongful act. For example, the defendant stabs and kills the victim before a fatal dose of poison previously administered by a third party can take effect. In duplicative causation cases, the effect of the defendant's act combines with, or duplicates, the effect of other potential causes where the latter were alone sufficient to bring about the injury. For example, the defendant and another party start separate fires that combine to burn down the victim's house where each fire was independently sufficient to do so. Since in these cases it is not true that but for the defendant's wrongful act the specified harm would not have occurred, according to the but-for test, in neither scenario is the defendant's conduct an actual cause of the injury. Such a result is contrary to intuitions about responsibility and, by implication, about causality. Clearly, in both scenarios, the defendant's act contributed to the injury.

To avoid these counterintuitive and unjust conclusions courts have, among a number of often *ad hoc* conceptual devices, generally relied on a *substantial factor* or *material contribution* test (see Wright, 1985, pp. 1777-84; Robertson, 1997, pp. 1775 ff.).[1] This test requires that the defendant's act must have been a substantial factor or materially contributed to the injury complained of. As Wright (1985, p. 1782) points out, this provides no test of causation but merely confounds the factual issue of whether the defendant's act was an actual cause of (was a factor in or contributed to) the specified injury and the *legal* issue of whether it was (material or substantial) enough of a factor for liability (legal responsibility) to attach.

1.2 The NESS Test

In an influential paper (Wright, 1985), Wright proposes what he describes as a comprehensive test for actual causation, the NESS *(N*ecessary *E*lement of a *S*ufficient *S*et) test: "a particular condition was a cause of (condition contributing to) a specific consequence if and only if it was a necessary element of a set of antecedent actual conditions that was sufficient for the occurrence of the consequence" (Wright, 1985, p. 1790).[2] Wright (1988, p. 1018) adopts the view that there is an intelligible, determinate concept of actual causation underlying and explaining common intuitions and judgements about causality and that this concept explains the "intuitively plausible factual causal determinations" of judges and juries when "not confined by incorrect tests or formulas." Wright (1985, p. 1902) contends that, not only does the NESS test capture

[1] In the case of Athey v. Leonati, [1996] 3 S.C.R. 458, the Supreme Court of Canada (per Major J.) confirmed these duo tests:

> The general, but not conclusive, test for causation is the "but for" test, which requires the plaintiff to show that the injury would not have occurred but for the negligence of the defendant....
>
> The "but for" test is unworkable in some circumstances, so the courts have recognized that causation is established where the defendant's negligence "materially contributed" to the occurrence of the injury.

[2] Wright's motivation is his concern that the absence of a comprehensive, workable definition of actual causation is exploited by certain "fashionable" jurisprudential camps (Libertarians, Legal Critics, Legal Economists, and Legal Realists) to undermine what Wright regards as the traditional moral basis of tort law, "a system of corrective justice based on individual autonomy and individual responsibility" (Wright, 1988, p. 1004).

the common-sense concept underlying these common intuitions and judgements, the NESS test defines the concept of actual causation.

Pearl (2000, pp. 313-15) claims that while the intuitions underlying the NESS test are correct the test itself is inadequate to capture these intuitions. He argues that the NESS test relies on the logical language of necessity and sufficiency and that traditional logic is incapable of capturing causal concepts. Pearl (Pearl, 1995; Galles and Pearl, 1997; Galles and Pearl, 1998; Pearl, 2000) proposes a mathematical language of (graphical) causal models employing structural equations for formalizing counterfactual and causal concepts. Pearl (1998; 2000, Chap. 10) first applies this structural language to define actual causation using a complex construction called a *causal beam*. (Halpern and Pearl, 2000) develops a "more transparent" definition, still using the language of structural models.

1.3 Thesis Statement

This thesis will investigate whether the definition of actual causation of Halpern and Pearl captures the meaning of Wright's NESS test and more generally whether the structural language developed by Pearl is adequate to formalize the NESS test. The thesis is: The Halpern and Pearl definition of actual causation fails to capture the meaning of the NESS test in the structural language of causal models. However, the alternative definition developed in this research does formalize the essential meaning of the NESS test in the structural language.

1.4 Thesis Overview

Chapter 2 presents and investigates the meaning of Wright's NESS test as a preliminary to investigating whether the Halpern and Pearl definition formalizes that meaning. Chapter 3 introduces Pearl's structural language and the Halpern and Pearl definition, and discusses the implications of a recent result of Hopkins and Pearl (2003) for the Halpern and Pearl definition. Chapter 4 develops an alternative structural definition of actual causation and counters arguments raised by Hopkins and Pearl against the possibility of defining actual causation in the structural language. Chapter 5 concludes the thesis with an investigation of the relation between the new definition of actual causation developed in this research and the NESS test.

2. The Origin And Meaning Of The Ness Test

As a necessary preliminary to exploring the possibility of a structural language formalization of the NESS test, this chapter explores the meaning of the NESS test by considering its origins, as described by Wright, and the meaning of its key concepts both through Wright's explanations and by comparison with other approaches, in particular, Mackie's concept of an "INUS condition" which is often identified with the NESS test.

The chapter is organized as follows: Section 2.1 outlines Wright's explanation of the origins of the NESS test. Section 2.2 compares the NESS test to the legally accepted test for actual causation, the counterfactual "but-for" test. Finally, Section 2.3 then considers the meaning that Wright attaches to the concepts that define the NESS test and problems inherent in his approach.

2.1 Wright on the Origins of NESS

According to Wright (1985, pp. 1789-91; 1988, pp. 1019-20; 2001, p. 1102), the NESS test is entailed by (what he describes as) the dominant regularity account of the meaning of general causation of Hume as modified by Mill. Mill argues that typically causation is a relation between a complex of multiple antecedent conditions and the consequent effect. An explosion that destroys a home might be explained as having been caused by the lighting of a match but the lighting of matches does not invariably result in explosions. However, if the presence of oxygen, the presence of a gas leak, and a certain concentration of gas (along with doubtless numerous other conditions) are added to the set of conditions then, on the regularity theory of causality, an explosion invariably follows. Wright (1985, p. 1790) interprets Hume as believing that in cases of causal regularity the effect not only inevitably occurs upon the occurrence of the cause, the effect only occurs upon that cause, "that a certain consequence is always produced by the same cause—that is, that there is a unique sufficient set of antecedent conditions that always must be present to produce a particular consequence." This is equivalent to

saying that a cause is both sufficient and necessary for its effect. Mill believes that for any given effect there are potentially many causes (the "plurality-of-potential-causes theory"; Wright 1985, p. 1790). The explosion of the house might have been caused, for example, in conjunction with the numerous other conditions, by the lighting of a match or by an electrostatic spark. Thus a particular effect may have been caused by one of several distinct but equally sufficient sets of conditions.

Wright attributes to Hart and Honore (1985, pp. 112-113) the idea of a contributing condition or *causally relevant factor* for some effect as any member necessary for the sufficiency of the jointly sufficient set of conditions for the effect (required by the regularity theory as modified by Mill), and thus attributes to Hart and Honore the origin of the NESS test (Wright 1988, p. 109):

> A particular condition was a cause of (contributed to) a specific result if and only if it was a necessary element of a set of antecedent actual conditions that was sufficient for the occurrence of the result.

For example, if the lighting of a match (and not an electrostatic spark) was a cause of the house explosion then the lighting of the match was a necessary member of the jointly sufficient set of actual (or realized) conditions that together caused that explosion; the jointly sufficient set of conditions of which an electrostatic spark is a member did not occur in the actual situation.

2.2 The NESS Test vs. the "but-for" Test

In circumstances where there is only one actual or potential set of conditions sufficient for the occurrence of the result, the NESS test reduces to the but-for test (Wright 1985, p. 1792; Wright 1988, p. 1021). While the but-for test is a necessity test, the NESS test subordinates necessity to sufficiency: necessity is in issue relative to an identified actually sufficient set of conditions (Wright 1988, p. 1019). Because of this, the NESS test can deal with those cases, involving multiple potential sets of sufficient conditions (overdetermination cases), where the but-for test failed (Wright 1988, p. 1021):

> A condition was a cause under the NESS test if it was necessary in the circumstances for the sufficiency of any actually sufficient set, even if, due to other actually or hypothetically sufficient sets, it was not—as required by the but-for test—necessary in the circumstances for the result.

(Another *actually* sufficient set would be present in a case of duplicative causation; a *hypothetically* or *potentially* sufficient set would be present in a case of preemptive causation—see Section 1.2.)

Wright (1988, p. 1020) argues that the choice of tests for actual causation should be governed by how well a test corresponds with common intuitions about the concept. To illustrate that the NESS test matches common intuitions where the but-for test fails he considers three variations of a two-fire scenario: fire *X* and fire *Y* are independently sufficient to destroy house *H* if they reach it and they are the only potential causes of house *H*'s destruction so that if neither reach the house it will not be destroyed. In the first variation, fire *X* reaches and destroys house *H* and fire *Y* would not have reached house *H* even if fire *X* were absent. The common intuition here is that fire *X* was a cause of the destruction of house *H* but not fire *Y*. In this case there is a single actually sufficient set of conditions and no other even potentially sufficient set of conditions. (This assumes that actually sufficient sets of conditions are minimal; that is, every element of the set is necessary for the sufficiency of the set.) Fire *X* was a necessary element (necessary for the sufficiency) of that single, actually sufficient set, a NESS condition. It was also a but-for condition.

In the second variation, fire *X* and fire *Y* reach house *H* simultaneously and destroy it together. Here Wright claims that the common intuition is that both (individually) fire *X* and fire *Y* were causes of the destruction of the house. There are two overlapping sets of actually sufficient conditions.[3] Fire *X* is necessary for the sufficiency of the set including itself but not fire *Y* and fire *Y* is necessary for the sufficiency of the set including itself buy not fire *X*. Neither fire *X* nor fire *Y* is a but-for cause of the destruction of house *H* but each is a *duplicative* NESS cause of the destruction.

In the final variation, fire *X* reaches and destroys house *H* before fire *Y* can arrive and, if fire *X* had been absent, fire *Y* would have destroyed house *H*. Here intuition

[3] Actually, there are at least three actually sufficient sets since the set including fire *X* would not cease to be sufficient by the addition of fire *Y*. That addition would result in neither fire *X* nor fire *Y* being necessary for the sufficiency of the set containing both. The problem is easily avoided by requiring that actually sufficient sets be minimal. However, the NESS test requires only that a cause (or contributing factor) be a necessary element of some, not every, actually sufficient set.

suggests that it is fire *X* that caused the destruction of house *H* and fire *Y* did not. Fire *Y* is not a NESS condition for the destruction of house *H* since any actually sufficient set of conditions, given the assumptions of the scenario, must include fire *X* and fire *Y* is not necessary for the sufficiency of any set of conditions that includes fire *X*. Fire *X*, on the other hand, is necessary for the sufficiency of the actually sufficient set of which it is a member. Because the set containing fire *Y* but not fire *X* would have been sufficient in the absence of fire *X*, fire *X* is not a but-for cause of the destruction of house *H*. Fire *X* was a *preemtpive* NESS cause because it preempted the actual sufficiency of the potentially sufficient set including fire *Y*.

2.3 The Meaning of NESS

Before considering whether and how the NESS test can be defined in the structural language and whether, in particular, the Halpern-Pearl definition of actual causation formalizes the NESS test it is necessary to attempt to understand the meaning Wright attaches to the key concepts of the test, the meaning of "necessary," "sufficiency," and "actually sufficient" in "necessary for the sufficiency of an actually sufficient set."

2.3.1 Necessary and Sufficient Conditions

Necessary and sufficient conditions are usually understood in terms of conditional statements. Conditional statements have the form "If *A* then *B*," meaning that if *A* is true (is the case, occurs, obtains) then B is true. *A* is called the *antecedent* and *B* is called the *consequent*. In traditional propositional logic, if-then statements are *material* conditionals and they mean that it is never the case that *A* is true and *B* is false. A true material conditional defines sufficiency and necessity relations between *A* and *B*: *A* is sufficient for *B* since whenever *A* is the case, *B* must be the case. (Since *B* can be true and *A* false, *B* is not sufficient for *A*). *B* is necessary for *A* since if *B* is not the case then *A* cannot be the case. (Since *A* can be false while *B* is true, *A* is not necessary for *B*).

Material necessity and sufficiency are exceedingly weak. Material conditionals operate on truth values and can be true when there is no connection (or, at least, no obvious connection) between the antecedent and consequent. For example, "If the United States is over 200 years old then the capital city of Canada is north of the capital city of Mexico" is a true material conditional. Worse, a false antecedent is materially

sufficient for any consequent and a true consequent is materially necessary for any antecedent. Material necessity and sufficiency, while they correctly handle relationships between truth values, have therefore not been thought serious candidates for explicating causal claims.

2.3.2 Counterfactuals

Counterfactual conditionals have played an important part in the analysis of causation. Counterfactual conditionals are a type of subjunctive conditional, conditionals that state what would happen (what the consequent would be) given a hypothetical (or modal) antecedent—if *A* were the case then *B* would be the case. Subjunctives in the past tense that imply that the antecedent is false are known as counterfactuals. For example, "If the fire had not been started, the house would not have been destroyed" or "But for the fire being started, the house would not have been destroyed." The but-for test is a counterfactual necessity test.

However, the meaning of counterfactuals themselves is disputed. Hume seems to identify his regularity theory of causation with a counterfactual definition:

> An object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed. (Hume An Enquiry Concerning Human Understanding, Section VII, Part II, quoted in (Mackie 1974, p. 30))

This position is generally rejected. Not only is it inconsistent with Mill's doctrine of the plurality of causes (above), identifying regularities is an observational task while applying counterfactuals is mental task (see Pearl 2000, p. 232). Indeed, many empiricists argue that counterfactuals are meaningless or indeterminate because they cannot be verified given that they refer to unactualized possibilities (see e.g. Dawid 2000; Menzies 2001).

2.3.3 Possible-Worlds Semantics for Counterfactuals

The best-known theory of counterfactuals is that of Lewis (1993). Lewis' theory is based on the concept of similarity among possible worlds and relies on the primitive relation of *comparative similarity*, a (weak) ordering of worlds in which one world is *closer to actuality* than another if the first resembles our world (the actual world) more than the

second does, and our actual world is closest to actuality. Lewis defines a counterfactual operator "$\Box\!\!\rightarrow$" so that $A \Box\!\!\rightarrow B$ (read "if A had been the case then B would have been the case") is true just in case A is true in all of the closest possible worlds where B is true (B-worlds). According to this account, the counterfactual "If A were not the case, then B would not be the case" is true in the actual world when the closest non-A worlds are non-B-worlds. This is the sense in which A is necessary for B. A would be sufficient for B if the closest A-worlds were B-worlds.

2.3.4 Mackie's "Nomic-Inferential" Model and the INUS Condition

Wright (1985, p. 1806; 1988, pp. 1041-42) rejects the idea that the causal enquiry involves attempting to construct counterfactual possible worlds. Wright describes the causal enquiry as an attempt "to determine which causal generalizations have been instantiated in the actual world by the conditions that occurred on the particular occasion" (Wright, 1988, p. 1042). In this respect Wright's approach to counterfactuals is most similar to that of Mackie (1993), the latter of which is dubbed the *nomic-inferential model* by Kim (1993).

Mackie (1974) follows Hume in distinguishing between our concept of causation (the concept supposedly underlying common causal intuitions or judgements) and causation in the objects by which he means relations in the objects that might justify or explain our concept. Mackie (1974, p. 57) argues that our concept of causation is inherently counterfactual:

> The key item is a picture of what *would* have happened if things had been otherwise, and this is borrowed from some experience where things *were* otherwise. It is a contrast case rather than the repetition of like instances that contributes most to our primitive concept of causation.

According to Mackie (1974, p. 31 ff.), our counterfactual thinking occurs relative to a background of circumstances, which is why, for example, we can say that when a struck match alights, a flame would not have occurred if the match had not been struck when we know it would have occurred if the match was touched by a red-hot poker. X is *necessary* for Y when X and Y are distinct, X and Y both occur, and, in the circumstances, if X had not occurred, then Y would not have occurred. X is *weakly sufficient* for Y when,

in the circumstances, if *X* occurs *Y* occurs. *X* is *strongly sufficient* for *Y* when, in the circumstances, if *Y* had not been going to occur, *X* would not have occurred.

Mackie (1974, pp. 53-54) argues that *contrary-to-fact conditionals* (or counterfactuals)—conditionals whose antecedent is "unfulfilled"—cannot be true or false. Counterfactuals are disguised arguments that rely for their validity on universal laws. For example, "If the match had not been struck the flame would not have appeared" represents the argument "The match was not struck so the flame did not appear" with an unstated premise to the effect matches never alight in such and such circumstances. This disguised-argument theory is what Kim calls the nomic-inferential model.

Mackie believes that, while strong sufficiency plays a part in our concept of causation—so that generally we will recognize *X* as a cause of *Y* if *X* is strongly sufficient for *Y*[4]—our concept requires only that *X* was necessary for *Y*. In other words, he accepts that the but-for test is consistent with our concept of actual causation. In support of this position he posits an indeterministic chocolate-bar-dispensing machine, *M*, which invariably dispenses a bar on insertion of an appropriate coin (an English shilling) but sometimes, indeterministically, does so without the insertion of any coin. By the definition of *M*, the insertion of a coin is not necessary for the production of a bar but, since in any circumstances the insertion of a coin in *M* produces a chocolate bar, is weakly sufficient for that result. Also, by definition of *M*, if a bar was not going to be produced, the insertion of a coin could not have been going to occur: the insertion of a coin (a shilling, of course) is strongly sufficient for the production of a bar. Because of the indeterministic nature of *M*, we cannot know on any particular occasion when a coin is inserted and the bar produced whether it would have happened even if the coin were not inserted. However, according to Mackie (1974, pp. 42-43),

> It is just this question that we need to have answered before we can say whether the insertion of the shilling caused the result. If the chocolate would not have come out if the shilling had not been put in, then the

[4] Note, however, if *X* is strongly sufficient for *Y* then *Y* is necessary for *X* in circumstances in which they both occur suggesting that if *X* is a cause of *Y* then *Y* is a cause of *X*. Mackie (1974, p. 51) gets around this arguing that our concept of causation also includes the relation of *causal priority*.

> insertion of the shilling caused the result. But if it would have come out anyway, the insertion of the shilling did not cause this. (This last ruling prejudges a question about causal over-determination that has still to be considered; but we shall reach an answer to this question which agrees with the present ruling.[5]) And, consequently, if it is in principle undecidable whether the chocolate would on this particular occasion have come out if the shilling had not been put in, it is equally undecidable whether the putting in of the shilling caused the appearance of the chocolate.

Machine *M* is contrasted with a second hypothesized, inderministic chocolate-bar-dispensing machine, *L*, which only dispenses a bar when an appropriate coin (a shilling) is inserted but sometimes, indeterministically, does not. In any case where candy is produced, the insertion of the coin was necessary and weakly sufficient but not strongly sufficient. The bar not being produced does not mean a coin could not have been inserted. Mackie argues that, nevertheless, our ordinary causal concepts require us to agree that inserting the coin causes the bar to be produced when it is produced, even though sometimes it fails to cause it to happen. Contrasting machines *L* and *M*, he concludes that *X* caused *Y* requires that, in the circumstances, *X* was necessary for *Y* and (trivially) weakly sufficient for *Y*, but not strongly sufficient for *Y*.

When Mackie turns from describing what he believes to be our ordinary concept of causation to causation "as it really exists in the objects" (Mackie, 1974, p. 59) he develops the concept of a cause as an INUS condition (Mackie, 1974, pp. 61-62): an insufficient but non-redundant part of an unnecessary but sufficient condition. Like the NESS test, Mackie's concept of an INUS condition is based on the regularity theory of general causation as modified by Mill's doctrine of the plurality of causes (Mackie 1974, pp. 61-62). According to Mackie, a causal regularity would have the form "All *ABC*[6] are followed by *P*" where *A*, *B*, *C*, and *P* are generalized types of events rather than specific

[5]Mackie deals with preemptive causation by arguing that our concept of causation requires that a cause be necessary for a result "as it came about" (Mackie, 1974, p. 46) and deals with duplicative causation by claiming that none of the duplicate causes are causes in themselves, but the aggregate of the duplicate causes satisfies the but-for test (Mackie, 1974, p. 47). Wright (1985, p. 1777; 1988, p. 1025) dismisses the first argument as proof by tautology and claims (Wright, 1985, p. 1777; 1988 p. 1027) the second argument allows irrelevant factors to be treated as causes.

[6] This is a conjunction of events $A \wedge B \wedge C$.

occurrences. If also "All *DGH* are followed by *P*" and "All *JKL* are followed by *P*" and *ABC*, *DGH*, and *JKL* are the only minimal sets of conditions that produce *P*—that is, assuming that a given event has a finitely-many distinct sets of conditions that produce it—and if *P* occurs only when one of *ABC*, *DGH*, and *JKL* occur—that is, every event has some set of conditions that produces it (every event has a cause)—then the disjunction of conjunctions $(ABC \vee DGH \vee JKL)$ is a necessary and sufficient condition for *P*. Respecting *ABC*, for example,

> *ABC* is a *minimal* sufficient condition: none of its conjuncts is redundant: no part of it, such as *AB*, is itself sufficient for *P*. But each single factor, such as *A*, is neither a necessary nor a sufficient condition for *P*. Yet it is clearly related to *P* in an important way: it is an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition (Mackie, 1974, p. 62).

As Kim (1993) summarizes, Mackie's position is that

> Singular causal assertions are explained in terms of the notion of "at least an INUS condition"; a cause of an event is at least an INUS condition of it. The notion of INUS condition in turn is explained on the basis of "necessary condition" and "sufficient condition", and these are analysed in terms of counterfactual conditionals. Finally, counterfactuals are explained on the nomic-inferential model. It is at this point that laws and regularities enter into singular causal judgements; according to Mackie, his analysis can be characterized as form of regularity theory of causation.

2.3.5 NESS vs. INUS

According to Wright (1988, p. 1041),

> The only question in the causal inquiry is whether the condition being tested was necessary on the particular occasion for the sufficiency of a set of actual antecedent conditions that was sufficient for the occurrence of the result—as required by the NESS test. There is an obvious, straightforward way to resolve this question. We hypothetically eliminate only the condition being tested…from the sufficient set of actual conditions. Then without adding or subtracting any other conditions, we determine—by matching the remaining conditions in the set against the applicable causal generalization—whether the set still would be sufficient for the occurrence of the result.

Wright (1985, 1988, and 2001) describes a causal law as an "if-then" statement the antecedent of which lists together "all of the conditions that are together sufficient for

the occurrence of the consequent" (2001, p. 1102). Causal generalizations are "incompletely specified causal laws that list only some of the NESS conditions along with the result, but nevertheless assert that the listed NESS conditions combine with unlisted, unknown NESS conditions to form a set of conditions that is sufficient for the result" (Wright 1988, p. 1031).

On the surface, it appears that the INUS and NESS accounts of actual causation are consistent. The description of the sense in which a condition is necessary for an effect is consistent with the nomic-inferential model. Also, according to Mackie's analysis, x is an actual cause of y, where x is an event of type A and y is an event of type P when (i) A is an INUS condition for P, (ii) all other members of at least one sufficient conjunct for P to which A belongs are present in the scenario in question, and (iii) at least one conjunct of any disjunct not containing A is absent from the scenario (Kim 1993). For example, if $(ABC \vee DGH \vee JKL)$ is a necessary and sufficient condition for P, then (i) is satisfied, (ii) is satisfied if B and C are present in the scenario, and (iii) is satisfied if at least one element from both $\{D, G, H\}$ and $\{J, K, L\}$ are absent from the scenario. However, if the disjuncts in $(ABC \vee DGH \vee JKL)$ were equated with the sufficient sets of Wright's NESS test, condition (iii) would rule out cases of duplicative causation (see Section 1.1). Wright (2001, p. 1130) calls a set of conditions satisfying conditions (i) and (ii) "analytically and empirically sufficient" but not "causally sufficient." To determine causal sufficiency Wright replaces condition (iii) with what he calls the "omnibus negative condition" which he again derives from Mill:

> [The failure of a sentry to be at his post] was no producing cause [of the army's being surprised by the enemy], but the mere absence of a preventing cause: it was simply equivalent to his non-existence. From nothing, from a mere negation, no consequences can proceed. All effects are connected, by the law of causation, with some set of positive conditions; negative ones, it is true, being almost always required in addition. In other words, every fact or phenomenon which has a beginning invariably arises when some certain combination of positive facts exists, provided certain other positive facts do not exist. . . .
>
> The cause then, philosophically speaking, is the sum total of conditions positive and negative taken together; the whole of the contingencies of every description, which being realised, the consequent invariably follows. The negative conditions, however, of any phenomenon, a special

> enumeration of which would generally be very prolix, may be all summed up under one head, namely, the absence of preventing or counteracting causes. (Mill A System of Logic, Bk. III, Ch. V, § 3, quoted in (Wright 2001, p. 1129))

(It is difficult to understand Mill's distinction; a method by which the enemy army might seek to produce surprise would be to eliminate the sentry.) The *negative conditions* or "absence of preventing or counteracting causes" (the absence of preemption) is what Wright defines as the omnibus negative condition; *positive conditions* are the set of analytically and empirically sufficient conditions.

Wright's (2001) account also differs from Mackie's with respect to the "complexity" which Wright allows to causal laws or generalizations. Wright's concept allows for "causal priority" among the conditions within a causal generalization. This notion arises in the context of his discussion of a class of cases that proved problematic for the NESS test, the so-called double omission cases:

> Some of the most difficult overdetermined-causation cases, at least conceptually, are those involving multiple omissions, which usually involve failures to attempt to use missing or defective safety devices or failures to attempt to read or heed missing or defective instructions or warnings. (Wright 2001, pp. 1123-1124).

Wright (1985, p. 1801; 2001, p. 1124 ff.) considers in detail the case of *Saunders System Birmingham Co. v. Adams*[7] where a car rental company negligently failed to discover or repair bad brakes before renting a car out. The driver who rented the car then negligently failed to apply the brakes and struck a pedestrian. In general, courts have held that individuals who negligently fail to repair a device (or provide proper safeguards or warnings) are not responsible when (negligently) no attempt was made to use the device (or use the safeguards or observe the warnings). According to Wright (2001, p. 1124), the court's decisions reflect a "tacit understanding of empirical causation in such situations": not providing or repairing a device (or not providing proper safeguards or warnings) can have no causal effect when no attempt was or would have been made to use the device (or use the safeguard or observe the warning)—unless

[7] Saunders Sys. Birmingham Co. v. Adams, 117 So. 72 (Ala. 1928).

no attempt was made because it was known that the device was inoperative (or the safeguards or warnings were inadequate).

Wright's (1985, p. 1801) original NESS analysis (where *D* represents the driver, *C* represents the car rental company, and *P* represents the pedestrian) is as follows:

> It is clear that *D*'s negligence was a preemptive cause of *P*'s injury, and that *C*'s negligence did not contribute to the injury. *D*'s failure to try to use the brakes was necessary for the sufficiency of a set of actual antecedent conditions that did not include *C*'s failure to repair the brakes, and the sufficiency of this set was not affected by *C*'s failure to repair the brakes. A failure to try to use brakes will have a negative causal effect whether or not the brakes are defective. On the other hand, *C*'s failure to repair the brakes was not a necessary element of any set of antecedent actual conditions that was sufficient for the occurrence of the injury. Defective brakes will have an actual causal effect only if someone tries to use them, but that was not an actual condition here. The potential negative causal effect of *C*'s failure to repair the brakes was preempted by *D*'s failure to try to use them.

Notice that interchanging *C* and *D*'s negligent acts in this argument results in an apparently equally plausible argument for *C*'s negligence being a preemptive cause of *P*'s injury. According to Wright (2001, p.1125),

> At the time that I wrote this explanation, I was aware that it was too brief and cryptic, relied upon an insufficiently elaborated notion of causal sufficiency and "negative causal effect," and therefore could seemingly be reversed to support the opposite causal conclusions merely by switching the references to the two omissions. Nevertheless, I thought it roughly stated the correct analysis in very abbreviated form.

Wright (2001, p. 1129) argues that this argumentative symmetry exists only when

> The NESS test is viewed "mechanically" as requiring mere analytical or empirical sufficiency. But it is not true if the test is properly understood as incorporating a concept of causal sufficiency, which requires the complete instantiation of the potentially applicable causal generalization, and if proper attention is paid to the distinction between positive and negative causal effects and the need to take into account any causal priority within an applicable causal generalization when assessing negative rather than positive causal effects.

In the case of the driver failing to apply defective brakes, Wright says that the issue is the cause of the brakes not being operated, "the failure of a causal generalization for

braking," a negative causal effect; when that is the case causal priority becomes important in applying the NESS test (Wright 2001, pp. 1130-31):

> The failure of any causal generalization is logically or empirically guaranteed to occur if any one of the necessary positive conditions in the antecedent of the causal generalization is absent. Yet, the failure can be explained causally only by taking into account any relevant causal priority among those positive conditions. The absence of any causally prior necessary condition preempts the possible coming into play (through presence or absence) of any other necessary condition in the causal generalization, the operation of which was causally subsequent to or dependent upon the causally prior necessary condition.

It is not easy to see, and Wright does not explain, how a non-structured set of conditions, related to each other only as instantiated antecedent conditions of some causal generalization, can have an internal causal structure. There is nothing in Mackie's account of causal laws that suggests the existence of causal priority among the elements of the minimal sufficient conditions or between sets of minimal sufficient conditions, rather, the structure of the relation between P and $(ABC \vee DGH \vee JKL)$ is logical identity: P *iff* $(ABC \vee DGH \vee JKL)$ (see Kim 1993).

"Causal priority" is not the only sense in which, it appears, that a causal law or generalization can have a causal structure, as evidenced by Wright's (1988, p. 1025) identification of actually sufficient sets with causal stories or chains in discussing Mackie's (1974, p. 44) well-known preemptive causation example of the ill-fated desert traveller:

> A desert traveller has two enemies intent upon his death. Enemy *A* poisons the traveller's water can. Enemy *B*, unaware of what enemy *A* has done, empties the traveller's water can before the traveller can drink from it. As a result, the traveller dies of dehydration rather than from poisoning.

Mackie's purpose in considering this example is to reconcile his contention that our concept of actual causation requires a cause to be a necessary (but-for) condition of the result (see Section 2.3.4) with the (supposed) common intuition that the emptying of the can caused the travellers death and not the poisoning of its contents, though neither satisfies the but-for test. Mackie (1974, pp. 45-46) argues that we recognize that the emptying of the can caused the traveller's death because we can complete the causal

story (or causal chain) involving the emptying of the can and death by dehydration but cannot complete the alternative story involving poisoning. According to Wright (1988, p. 1025) this account of why we recognize the emptying of the can as the cause of the traveller's death has merit only because it implicitly invokes the NESS test: a "completed causal story or causal chain is simply an actually sufficient set."

2.3.6 A Language for NESS

The idea that a single causal generalization can exist for a causal chain of events raises the issue of how the causal generalizations for the individual links in the chain relate to the causal generalization for the chain as a whole. According to Wright (1988, p. 1042), in applying the NESS test,

> We are trying to determine which causal generalizations have been instantiated in the actual world by the conditions that occurred on the particular occasion. Thus, we do not change any causal generalization. Nor do we need to worry about changing the prior conditions that produced the condition being tested [for actual causation]. The effects of these prior conditions are incorporated in the particular sufficient set of existing conditions, which is a time-slice view of the ongoing causal network. When we hypothetically eliminate the condition being tested we automatically hypothetically eliminate the effects of prior conditions insofar as they operate through the condition being tested.

Suppose A is a necessary member of $\{A, A_1, ..., A_n\}$, an actually sufficient set of conditions for a result B (the "time-slice view" from A to B) and that B is a necessary member of $\{B, B_1, ..., B_m\}$, an actually sufficient set of conditions for a result C. The expectation is that A is an actual cause of C (e.g., A: X shoots a rope holding a piano being lowered from an apartment window; B: the rope holding the piano being lowered from the apartment window breaks; and C: the piano falls to the ground injuring a passing pedestrian). If $\Psi = \{A, A_1, ..., A_n, B, B_1, ..., B_m\}$, then Ψ should be an actually sufficient set of conditions for C (in the time-slice view from A to C). However, it does not seem possible for both elements of $\{A, A_1, ..., A_n\}$ and B to be necessary elements of Ψ. If this intuition is correct, then relative to the causal generalization for the entire chain, either A or B is not a cause of C.

According to Pearl (2000), this problem is symptomatic of attempting to explicate causal intuitions in the language of logical necessity and sufficiency. He describes it as "syntax sensitivity." If, for instance, $A \; iff \; B \vee C$ and $D \; iff \; A \vee E$ then $D \; iff \; B \vee C \vee E$. Should we conclude, asks Pearl, that A is not a cause of D given $D \; iff \; B \vee C \vee E$? Pearl (2000, p. 315) argues that the "structural information conveying the flow of influences in the story cannot be encoded in standard logical syntax."

Another limitation with a logical account, raised by Pearl, is the logical equivalence of "if *A* then *B*" and "if $\neg B$ (not-*B*) then $\neg A$," an inversion "not supported by causal implications." According to Wright (2001, p. 1103 n. 113), however,

> I have always viewed the NESS test as embodying not merely a requirement of logical or even empirical necessity or sufficiency, but also a notion of causal directionality according to which the conditions specified in the antecedent ("if" part) of the causal generalization are causally relevant conditions for the occurrence of the condition specified in the consequent ("then" part), but not vice versa, and a notion of causal sufficiency which requires that all the conditions specified in the antecedent and the consequent be concretely instantiated on the particular occasion.

Pearl (2000, Ch. 10) argues that his "structural logic" language (see Sections 3.2 to 3.4) is adequate to capture and formalize the intuitions that underlie the INUS condition, which he (mistakenly, according to the arguments of the previous section) identifies with the NESS test. The remainder of the thesis will consider whether the NESS test can be formalized in Pearl's structural language.

2.3.7 The Problem of Circularity

Fumerton and Kress (2001, p. 102) argue that

> Wright's project is to analyze the meaning of the word or, alternatively, the concept of causation. If he deploys the concept of a *causal* law in defining causation, surely his critics will charge him with a vicious form of circularity—his NESS test for causation is nearly tantamount to defining causation as causation.

However, as explained above, Wright argues (in effect) that the NESS test is a corollary of the regularity theory of causation. Wright would be open to the charge of "vicious circularity" if he argued that the NESS test identifies causal laws or regularities, but he

does not. The NESS test is concerned with what it is to attribute causation to some factor in a specific scenario using existing causal knowledge which, he argues (see e.g. Wright, 1988, p. 1045), contrary to Mackie, is represented as causal laws or generalizations.

In summary, then, the NESS test requires the existence of an existing set of conditions, including the putative causal condition, that in some sense guarantees the effect in question ("actual sufficiency") but does not guarantee the effect if the putative cause is removed from the set ("necessary for the sufficiency"). These sufficient sets are recognized because of existing causal knowledge that is represented as causal laws. A logical interpretation of these requirements allows for transformations and substitutions that can "break" the test. A language for formalizing the NESS test needs to prevent such operations and allow for causal and temporal priority relationship among the conditions in a sufficient set. The next chapter introduces the Pearl's structural language, which appears to satisfy these criteria.

3. Structural Definition Of Actual Causation

In broad outline, Pearl's "structural language" response to Wright's account of the NESS test and the problems that arise with that account might be as follows. Causal generalizations can have a causal structure because they are relative to (informal) causal models built up from internal (mental) representations of causal mechanisms, that is, invariant relationships between a variable representing an event (the effect) and one or more variables representing other events (the causes or causal factors)—invariant under interventions that manipulate the value of the causal variables. Causal mechanisms link together into causal models via shared variables; that is how causal generalizations for individual events can link into extended causal chains or stories. Causal mechanisms describe functional, one-way (non-equational) relationships between the dependent (effect) variable and the independent (causal) variables; that explains why causal generalizations are not reversible. Counterfactuals describe the response of the value of variables in the model to interventions fixing the value of some other variables. An actual causal query is fundamentally a counterfactual query relative to some (permissible) set of interventions fixing the value of some of the model variables (to their actual values or otherwise).

This chapter elaborates on the preceding broad outline as follows. Section 3.1 describes the manipulability account of causation to which Pearl's structural language belongs. Sections 3.2 to 3.5 outline the structural language that is the basis of the Halpern-Pearl definition of actual causation presented in Section 3.7. Finally, Section 3.8 considers whether, in light of recent work by Hopkins and Pearl (2003), the Halpern-Pearl definition is adequate to formalize the NESS test.

3.1 Manipulability Theories of Causation

As Hausman and Woodward (1999, p. 533) write,

> One crucial fact about causation, which is deeply embedded in both ordinary thinking and in methods of scientific inquiry, is that causes are as it were levers that can be used to manipulate their effects. If *X* causes *Y*, one can wiggle *Y* by wiggling *X*, while when one wiggles *Y*, *X* remains unchanged.... For most scientists, the crucial difference between the claim that *X* and *Y* are [merely] correlated and the claim that *X* causes *Y* is that the causal claim, unlike the claim about correlation, tells one what would happen if one intervened and changed the value of *X*. It is this feature of causal knowledge that is so important to action.

The concept of causes as levers for manipulating effects is known as a "manipulability theory of causation" (see Woodward 2001). Different flavours of manipulability theories are characterized by their understanding of the nature of the manipulations that found the causal claims, whether manipulation is itself a causal notion. Traditional philosophical manipulability theories attempt to treat manipulations as primitive, non-causal concepts or define them in terms of other supposedly non-causal concepts such as human agency or acts of free will (see Woodward 1999, 2001). Non-reductionist theories accept that manipulations are inherently causal. That leaves them vulnerable to analogous critiques to that of Fumerton and Kress with respect to Wright's NESS test, that the account is circular (see Section 2.3.7). Two responses are, first, that the concept of manipulation is used to characterized a particular relationship as causal or not without any prior information about that relationship (i.e., informative if non-reductive) and, second, there is more than one notion of causation and the causal concept assumed in the notion of manipulation can be used to identify related concepts (Woodward 2001).

Both of these responses are evident in the work of Judea Pearl, whose non-reductionist manipulability theory relies on the "structural" notion of an "intervention" and the related concept of causal mechanisms. Much of Pearl's work is dedicated to explaining how causal relationships can be discovered from raw statistical data and from raw statistical data combined with substantive causal assumptions (see Pearl 2000, Chapters 1-3). Equally important are issues of what can be done with causal knowledge, what inferences or predictions can be made; and what is done, how causal knowledge can justify, explain, and explicate causal thought and behaviour including counterfactual and actual causal assertions. Key to both aspects are interventions and causal mechanisms.

3.2 Causal Mechanisms

According to Pearl (2000, p. 223),

> The world consists of a huge number of autonomous and invariant linkages or mechanisms, each corresponding to a physical process that constrains the behavior of a relatively small group of variables

The invariant relationships that are causal mechanisms can range from visible and tangible (e.g., a car's brake mechanism) to conceptual or theoretical based on physical, chemical, or social laws and conventions such as legal rules (Drudzel and Simon 1993); for an in-depth discussion of causal mechanisms and a comparison with traditional accounts of causal laws see (Woodward 2000, 2002).

When mechanisms link together (by shared variables) into groups, the *invariance* property of causal mechanisms means that changes (interventions) to one mechanism do not change the other mechanisms in the group. That is not to say that disturbing a mechanism will not change the values of variables related by other mechanisms, but the relationship between the variables, the relationship that defines the causal mechanism, remains unaltered. Because of this *autonomy* property of causal mechanisms, according to Pearl (1999, p. 1445), identifying causal mechanisms amounts to acquiring "knowledge" since it identifies patterns of behaviour (regularities) transportable across different situations. The *modularity* (see Pearl 2000, p. 22) of causal knowledge encoded in causal mechanisms explains our familiarity with, and our ability to comprehend, causal relationships used in causal explanations of novel situations (see Pearl 2000, p. 26; Pearl 1999, p. 1445). Causal mechanisms also explain why we readily teach each other what the normal results of actions are and why we readily predict the consequences of most actions: actions can be represented as *local surgeries* in the space of causal mechanisms. Actions are disturbances in the space of mechanisms. If there is a common understanding of how some group of mechanisms interact (or link) with each other for some part of the world then the effect of an action is understood as the new equilibrium reached by the modified group of mechanisms after the few mechanisms

disturbed by the action are respecified (see Pearl 2000, p. 223). *Locality* refers to an action only disturbing a few mechanisms (because of autonomy), for example,

> Tipping the leftmost tile in an array of domino tiles does not appear to be "local" in physical space, yet it is quite local in the mechanism domain: only one mechanism is perturbed, the gravitational restoring force that normally keeps that tile in a stable erect position; all other mechanisms remain unaltered, as specified, obedient to the usual equations of physics. Locality makes it easy to specify this action without enumerating all its ramifications. The listener, assuming she shares our understanding of domino physics, can figure out for herself the ramifications of this action, or any action of the type: "tip the *i*th domino tile to the right." By representing the domain in the form of an assembly of stable mechanisms, we have in fact created an oracle capable of answering queries about the effects of a huge set of actions and action combinations—without us having to explicate those effects. (Pearl 2000, p. 224).

Causal assertions arise as abbreviations about events related by causal mechanisms. Where mechanisms are unnamed, as most are in ordinary discourse, actions may be characterized by their immediate effects, as in, "the left-most tile is tipped to the right." These abbreviations suffice when there is a common understanding of the domain knowledge (the relevant group of linked mechanisms) in that it should be possible to determine what mechanism must be perturbed to bring about the specified event and what the other consequences of that would be, for example,

> This linguistic abbreviation defines a new relation among events, a relation we normally call "causation": Event *A* causes *B* if the perturbation needed for realizing *A* entails the realization of *B*. Causal abbreviations of this sort are used very effectively for specifying domain knowledge. Complex descriptions of what relationships are stable and how mechanisms interact with one another are rarely communicated explicitly in terms of mechanisms. Instead, they are communicated in terms of cause-effect relationships between events or variables. We say, for example: "If tile *I* is tipped to the right, it causes tile $i + 1$ to tip to the right as well"; we do not communicate such knowledge in terms of the tendencies of each domino tile to maintain its physical shape, to respond to gravitational pull and to obey Newtonian mechanics. (Pearl 2000, p. 225-226)

3.3 From Causal Mechanisms to Causal Models

The notion of autonomous causal mechanism explains the existence of *systems*, parts of the world that may be studied in isolation even though they are still linked to the rest of the world; for example, natural systems such as a hurricane or an eye, or artificial systems like a computer. Simon (1969) argued that the elements within systems are strongly interconnected while the connections to the outside world are relatively weak.

The abstraction of a system is called a *model*, a simplified representation that makes it possible to study how features of the system regarded as important interact. Models can range in complexity from informal, mental models to formal, mathematical models. Often, scientists represent models of systems as sets of simultaneous, algebraic equations. Pearl argues that algebraic equations are inadequate to represent systems of causal mechanisms and gives as an (archetypal) example a model for an artificial system, a circuit diagram for some electric circuit (Pearl 2000, pp. 346-47). The gates in the diagram represent (physical) causal mechanisms—they are autonomous (or independent) and invariant; perturbing or changing one gate does not affect the others and they behave the same in one circuit as they would in another. The diagram enables us to predict not only how the circuit will behave under normal conditions (an unperturbed domain model; i.e., in equilibrium) but also how the circuit will behave under abnormal conditions (perturbations of one or more gates—i.e., mechanisms—deliberately or otherwise). As Pearl (2000, p. 344) explains,

> For example, given [a] circuit diagram, we can easily tell what the output will be if some input changes from 0 to 1. This is normal and can easily be expressed by a simple input-output equation. Now comes the abnormal part. We can also tell what the output will be when we set [the input of some internal gate] to 0 (zero), or tie it to [the value of some external input], or change this *and* gate to an *or* gate, or when we perform any of the millions of combinations of these operations.

However, the equations describing the input-output conditions for a given gate do not allow us to predict the outcome of abnormal occurrences unguided by the diagram. Consider the system in Figure 3.1 consisting of only a multiplier (*2) and an adder (+1) (Pearl 2000, pp 346-7).

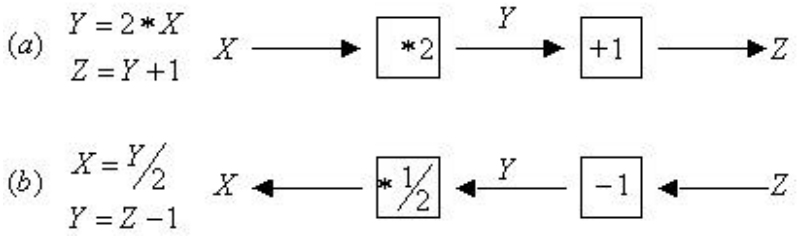

Figure 3.1: Unperturbed Circuit Diagram Model

The equations on the left in (a) describe the input/output conditions of the gates but are not equivalent to the corresponding diagram: we can algebraically manipulate the equation into the equivalent form (b) which corresponds to a diagram representing a system of gates different from that in (a). In other words, the diagrams represent critical information about the (causal) nature of the electric circuit that the equations alone cannot capture. The diagram in (a) shows that if we physically manipulate (intervene, perturb, change) Y it will affect Z while in (b) it will affect X with no effect on Z. Figure 3.2 demonstrates the relation between manipulations on the diagrams (representing physical manipulations) and the corresponding equations: setting Y to 0 removes the link between X and Y. The new mechanism controlling Y is represented by the equation $Y = 0$ which replaces $Y = 2X$ in (a) and $Y = Z-1$ in (b). The result of the manipulation is the solution of the new set of equations, i.e., $Z = 1$ in (a) and $X = 0$ in (b).

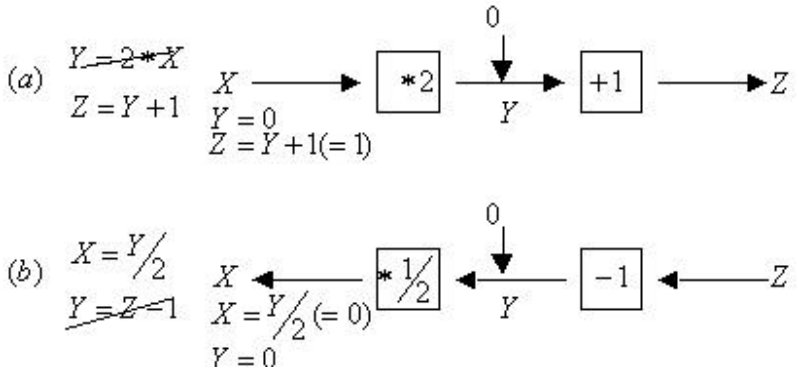

Figure 3.2: Perturbed Circuit Diagram Model

According to Pearl (2000, p. 347),

> We now see how this model of intervention leads to a formal definition of causation: 'Y is a cause of Z if we can change Z by manipulating Y,

> namely, if after surgically removing the equation for *Y*, the solution for *Z* will depend on the new value we substitute for *Y*". We also see how vital the diagram is in this process. *The diagram tells us which equation is to be deleted when we manipulate Y.* That information is totally washed out when we transform the equations into algebraically equivalent form.

Thus, interventions correspond to the "wiping out" of or "surgery" on equations guided by the diagram—and causation is a linguistic summary of the consequences of the surgeries.

An equation that represents an autonomous mechanism is a *structural equation.* (The identification of causal mechanisms with structural equations was made first by Simon (1953).) Pearl (2000, p. 160) gives the following (manipulationist) definition of a structural equation: An equation $y = bx + e$ is said to be *structural* if it is to be interpreted as follows. In an ideal experiment where we intervene to fix X to x and any other set Z of variables (not containing X or Y) to z, the value of Y is given by $bx + e$, where e is not a function of the settings $X = x$ and $Z = z$. The equality sign ("=") in a (parameterized) structural equation has a dual interpretation: it is symmetrical in relating the variables X and Y so that observing $Y = 0$ implies $bx = -e$; but it is asymmetrical with respect to interventions. In that case the equality sign stands for "is determined by" and setting Y = 0 gives no information about how x and e are related.[8] The nonparametric analogue has the form $Y = f(x,e)$ and represents Y as a non-specified function of X. In both cases, e represents an error term for omitted (non-modelled) factors or disturbance external to the system. In this case, Y is determined by only one observed variable, X. In the general case, $Y = f(pa_i, e)$ where pa_i represents the minimum set of variables under consideration such that if Z is any other set of variables

[8] As often happens in AI discussions of this problem (see Pearl 2000), the "=" sign is overloaded in this thesis. In addition to the dual relational/functional interpretation for structural equations described here, the equals sign may be used in logical equations to describe the truth conditions of a proposition, it may be used to represent logical equivalence between propositional statements, and it may be used to describe the value of a random variable (an event). The context of the discussion should make clear which usage is intended.

(distinct from Y and pa_i) the value of Y would be independent of z. (In other words, for a given e, if the values of the pa_i are fixed, then Y will be a constant or trivial function of the variables in Z.) The choice of the symbol "pa_i" is non-accidental since, graphically, they have a natural interpretation as the parents of Y and represent direct causes of Y.

A system of equations in which each equation represents an autonomous mechanism is a *structural model*. When each structural equation determines the value of a single, distinct variable (the *dependent* variable) the model is a *structural causal model*. While systems of algebraic equations jointly model some part of the world, any subset of a system of structural equations is itself a valid model of reality "that prevails under some set of interventions" (Pearl 2000, p. 27).

3.4 Structural Causal Models

A *causal model* (or *structural model*) models a system of causal mechanisms. The events (or event states) that causal mechanisms relate are represented as random variables; a particular value (a realization) of a variable X has the form $X = x$. (For example, if X is a variable for a light switch, $X = 1$ might represent the event that the light switch is "on", and $X = 0$ that it is "off".) The variables in a causal model are either *endogenous* or *exogenous*. The values of exogenous variables are determined by non-modelled factors; their values are taken as given, the model does not explain them.[9] The values of endogenous variables are determined entirely by the other variables (endogenous and/or exogenous) in the model. The relations that describe the values of endogenous variables as functions of the other variables are structural equations; that is, the system's causal mechanism are represented in the model as structural equations.

The following descriptions and definitions follow closely the corresponding expositions in (Halpern and Pearl 2000 and Pearl 2000):

Formally, a *signature* $\mathcal{S}$ is a 3-tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ is a relation associating

[9] Pearl (2000, p. 207) calls a causal model with a particular realization for the exogenous variables a causal *world* or *theory*.

with each variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (the set of values over which Y *ranges*; more typically referred to as the *domain* of *Y*, denoted $Dom(Y)$).

A *causal model* over a signature $\mathcal{S}$ is a 2-tuple $M = (\mathcal{S}, \mathcal{F})$, where $\mathcal{F}$ is a relation associating each $X \in \mathcal{V}$ with a function denoted F_X such that, $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V}-X} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$ (or $F_X : Dom(W) \rightarrow Dom(X)$, where $W = \mathcal{U} \cup (\mathcal{V} - X)$).

The function F_X for X is the structural equation describing the mechanism that determines the value of X (Typically, the F_X is expressed as an equation for X.)

Reflecting the assumption that an effect cannot precede its cause, the models considered here are *recursive* (or *acyclic*), meaning that there is a total ordering[10] $\prec$ of the variables in $\mathcal{V}$ such that if $X \prec Y$, then the value of F_X is independent of Y (i.e., $F_X(\ldots, y, \ldots) = F_X(\ldots, y', \ldots)$ for all $y, y' \in \mathcal{R}(Y)$). Recursive models have a unique solution for a given setting $\vec{u}$ of the variables in $\mathcal{U}$ (such a setting $\vec{u}$ is called a *context*). If PA_X is the minimal set of variables in $\mathcal{V}$ - X and U_X the minimal set of values in $\mathcal{U}$ that together suffice to represent F_X (i.e., if $Y \in \mathcal{V}$ ($Y \in \mathcal{U}$) and $Y \notin PA_X$ ($Y \notin U_X$) then F_X is independent of Y), then the causal model gives rise to a *causal diagram*, a directed acyclic graph (DAG) where each node corresponds to a variable in $\mathcal{V}$ and the directed edges point from members of PA_X and U_X toward X. The set PA_X, connoting the *parents* of X, are the *direct causes* of X. Causal diagrams encode the information that the value of a variable is independent of its other ancestor variables in the diagram given the values of its parents and, also, that the value of a variable can only affect the value of

[10] A total ordering is a relation R that is reflexive (xRx), transitive (xRy and yRz implies xRz), and asymmetric (xRy and yRx implies $x = y$), and such that for any two elements x and y, either xRy or yRx.

its descendents in the diagram. The edges in a causal diagram represent the non-parameterized (or arbitrary) form of the function for a variable, $X = F_X(U_X, PA_X)$.

An external *intervention* (or perturbation or surgery) setting $X = x$ (representing contingencies outside of the model perturbing a causal mechanism), where $X \in \mathcal{V}$, is denoted $do(X = x)$ or $X \leftarrow x$ and amounts to pruning the equation for X from the model and substituting $X = x$ in the remaining equations. In the corresponding causal diagram, it amounts to removing the edges from $PA_X \cup U_X$ to X. An intervention that forces the values of a subset of variables in $\mathcal{V}$ (represented sometimes by the ordered-set or vector notation, $\vec{X} \subseteq \mathcal{V}$: $do(\vec{X} = \vec{x})$ or $\vec{X} \leftarrow \vec{x}$) prunes a subset of equations, one for each variable in the set and substitutes the corresponding forced values in the remaining equations. That the resulting model is still a valid representation of reality "that prevails under some set of interventions" (Pearl 2000, p. 27).

Interventions map causal models into causal models. The model resulting from an intervention is a *submodel*: Given a causal model $M = (\mathcal{S}, \mathcal{F})$, a (possibly empty) subset $\vec{X}$ of variables in $\mathcal{V}$, and vectors $\vec{x}$ and $\vec{u}$ of values for $\vec{X}$ and $\mathcal{U}$ respectively, $M_{\vec{X} \leftarrow \vec{x}}$ denotes a new causal model over the signature $\mathcal{S}_{\vec{X}}$ $\mathcal{S}_{\vec{X}} = (\mathcal{U}, \mathcal{V} - \vec{X}, \mathcal{R}\big|_{\mathcal{V}-\vec{X}})$. $M_{\vec{X} \leftarrow \vec{x}}$ is called a submodel of M. Formally, $M_{\vec{X} \leftarrow \vec{x}} = (\mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X} \leftarrow \vec{x}})$ where, for $Y \in \mathcal{V} - \vec{X}$, $F_Y{}^{\vec{X} \leftarrow \vec{x}}$ is obtained from F_Y by setting the values of the variables in $\vec{X}$ to $\vec{x}$. The submodel $M_{\vec{X} \leftarrow \vec{x}}$ represents the *effect of action* $do(\vec{X} = \vec{x})$ on the model M.

3.5 Structural Definition of Counterfactuals

Pearl (2000, pp. 217-220) regards the word "counterfactual" as a misnomer to the extent it implies that counterfactual statements are contrary to facts or are not amenable to empirical verification. Pearl argues that counterfactuals are a "roundabout" way or "conversational shorthand" for stating predictions based on empirical causal knowledge.

Pearl gives the example of Ohm's law $V = IR$ the empirical content of which can, he argues, can be represented in a predictive or counterfactual form:

Predictive form: If the current at time t_0 is I_0, $I(t_0) = I_0$, then, with all else held equal, at any future time $t > t_0$

$$V(t) = \frac{V_0}{I_0} I(t).$$

Counterfactual form: If at time t_0 $I(t_0) = I_0$ and $V(t_0) = V_0$ then, had $I(t_0)$ been I' instead of I_0, the voltage would have been

$$V' = \frac{V_0 I'}{I_0}.$$

According to Pearl, both forms allow an infinite number of predictions from the single measurement (I_0, V_0) and both depend upon a scientific law ascribing a time-invariant property (the ratio V/I) to any object conducting electricity.

Pearl suggests two reasons why the counterfactual form is used. The first reason is to convey the logical consequences of a prediction. The intent of saying, "If you had left the lights on, the battery would be dead" may be to convey that the lights were not left on; that implicit assertion is explicitly justified by a logical implication (prediction) based on a general law. The second reason relates to the "all else held equal" requirement in the predictive form (above). In the case of predictive claims, what must be held equal needs to be carefully specified. On the other hand, many of these specifications are implicit (and do not need to be stated) in counterfactual expressions, especially when the underlying causal model is agreed upon.

If a counterfactual statement is to be interpreted as conveying a set of predictions then, according to Pearl, two components must remain invariant: the laws (or causal mechanisms) and the boundary conditions. In a causal model, these correspond to the functions F_X (for $X \in \mathcal{V}$) and the background variables $\mathcal{U}$. This means that the validity of the predictive interpretation of counterfactuals requires the assumption that variables

in $\mathcal{U}$ remain invariant when interventions in the model are made to represent the set of conditions in the prediction.

Formally, given sets of variables $\vec{X}$ and $\vec{Y}$ in $\mathcal{V}$ and a context $\vec{u}$, the *counterfactual* sentence, "The value that $\vec{Y}$ would have obtained, had $\vec{X}$ been $\vec{x}$" is interpreted as denoting the *potential response* for $\vec{Y}$ and $\vec{u}$ under the intervention $do(\vec{X} = \vec{x})$ (denotable as $F_Y(\vec{u},\vec{x})$), the solution for $\vec{Y}$ in the (sub)model $M_{\vec{X}\leftarrow\vec{x}}$ (see Pearl 2000, p. 204).

Pearl believes the reason why counterfactuals play an important part in causal explanations is that the utility of a causal explanation "is proven not over standard situations but rather over novel settings that require innovative manipulation of the standards" (Pearl 2000, p. 219). Submodels, which describe counterfactual worlds, result from the manipulation of causal mechanisms, whose autonomy is an "open invitation" to remove or replace them. The explanatory value of sentences is judged by how well they predict the ramifications of these interventions; that is, the validity of the counterfactuals they give rise to.

3.6 The Halpern-Pearl Structural Definition of Actual Causation

The following descriptions and definitions follow closely the exposition in (Halpern and Pearl 2000):

For a given signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$: A *primitive event* is a formula of the form $X = x$, where $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *basic causal formula* is of the form $[Y_1 \leftarrow y_1,..., Y_k \leftarrow y_k]j$ where j is a Boolean combination of primitive events,[11] $Y_1,...,Y_k$ are distinct variables in $\mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$. Basic causal formulas are abbreviated as $[\vec{Y} \leftarrow \vec{y}]j$ or just j when $k = 0$. A *causal formula* is a Boolean combination of basic causal formulas.

[11] A combination of primitive events produced from repeated application of the negation ("$\neg$") and conjuction ("$\wedge$") connectives.

A basic causal formula is true or false in a causal model given a context $\vec{u}$. Where y is a causal formula, $(M,\vec{u}) \models y$ means that y is true in the causal model M in the context $\vec{u}$. $(M,\vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X = x)$ means that the variable X has value x in the unique solution to the equations in the submodel $M_{\vec{Y} \leftarrow \vec{y}}$ in context $\vec{u}$. (In other words, in the counterfactual world $M_{\vec{Y} \leftarrow \vec{y}}$, resulting from the intervention $do(\vec{Y} = \vec{y})$, X has the value x.) When j is an arbitrary Boolean combination of primitive events $(M,\vec{u}) \models j$ is treated similarly; for example, $(M,\vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X_1 = x_1 \wedge \neg(X_2 = x_2))$ means that X_1 has the value x_1 and X_2 does not have the value x_2 in the unique solution to the equations in $M_{\vec{Y} \leftarrow \vec{y}}$ given context $\vec{u}$.

The types of events that count as causes are conjunctions of primitive events of the form $X_1 = x_1 \wedge ... \wedge X_k = x_k$, which may be abbreviated as $\vec{X} = \vec{x}$. (In practice, Halpern and Pearl considered only singular primitive events as causes; it was later proven that the definition requires this in the case of finite models. See below.) Events caused can be any Boolean combination of primitive events.

Definition (a*ctual cause; Halpern-Pearl version*): $\vec{X} = \vec{x}$ is an *actual cause* of j in a model *M* in the context $\vec{u}$ (i.e., in $(M,\vec{u})$) if the following conditions hold:

C1. $(M,\vec{u}) \models (\vec{X} = \vec{x}) \wedge j$.

C2. There exists a partition $(\vec{Z},\vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}',\vec{w}')$ of the variables in $(\vec{X},\vec{W})$ such that, where $(M,\vec{u}) \models Z = z^*$ for each $Z \in \vec{Z}$ [12],

(a) $(M,\vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}']\neg j$, and

(b) $(M,\vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]j$ for every subset $\vec{Z}'$ of $\vec{Z}$.

[12] In other words, for any variable *Z* in $\vec{Z}$ the actual value of *Z* in the model *M* given the setting $\vec{u}$ is represented as *z**.

C3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions C1 and C2.

Condition C1 is just the requirement that $\vec{X} = \vec{x}$ (the putative cause) and j (the effect whose cause is being investigated) be true in the original model given the setting $\vec{u}$ of the exogenous variables. Condition C2 has its origins in the concept of *sustenance* developed by Pearl (2000, Section 10.2) to represent what Hall (2003) identifies as the two concepts of causation, *dependence* and *production*. The dependence aspect of causation refers to the necessity (but-for) of a cause in maintaining an effect relative to interventions in the actual world. The production aspect refers to the capacity of a cause to bring about an effect in a context in which both are absent. Pearl conceives of the notion of sustenance as dependence "enriched" with features of production in a world (context) where both are true.

Condition C2(a) represents the dependence aspect according to which, where $\vec{X} = \vec{x}$ is a cause of j, if $\vec{X} \neq \vec{x}$ then j should be false ($\neg j$). However, unlike the traditional but-for test, C2(a) allows interventions in the model (structural contingencies) that make j false—the identification of a set of variables $\vec{W}$ and a setting $\vec{w}'$ that makes j false—when otherwise j would still be true despite setting $\vec{X} \neq \vec{x}$. Remembering the premptive causation cases, the motivation for allowing such interventions should be clear; for example, in the two-fire cases, where fire *X* arrives and destroys the house *H* before fire *Y* can arrive and do so (see Section 2.2), C2(a) allows for testing the dependence on the presence of fire *X* of the destruction of *H* under the contingency that fire *Y* is not present. If $\vec{X} = \vec{x}$ is a cause of j then necessarily there are some structural modifications of the model under which j is false when $\vec{X} \neq \vec{x}$.

Condition C2(b) represents the production aspect. If $\vec{X} = \vec{x}$ is an actual cause of j then when $\vec{X}$ is returned to its actual value under setting $\vec{u}$ then j should once again be true, both despite the structural changes made in C2(a) to make j false (the intervention $\vec{W} \leftarrow \vec{w}'$) and not as the result of those contingencies; that is, it is necessary to guard against j being true when $\vec{X}$ is returned to its actual value because of changes

in the values of the other variables in $\vec{Z}$ $(\vec{Z} - \vec{X})$. Restoring the values of the variables in any subset of $\vec{Z}$ to their original values in context $\vec{u}$ should not make j false. C2(b) ensures that $\vec{X} = \vec{x}$ alone is sufficient for maintaining j in context $\vec{u}$.

The variables in a set $\vec{Z}$ satisfying condition C2 mediate between $\vec{X}$ and j, and should be thought of as describing the "active causal process" from $\vec{X}$ to j. A minimal set $\vec{Z}$ satisfying condition C2 is defined by Halpern and Pearl as an *active causal process* and they show that every variable in an active causal process lies on a path from a variable in $\vec{X}$ to a variable in j and changes value when $(\vec{X}, \vec{W})$ is set to $(\vec{x}, \vec{w})$ in C2.

According to Halpern and Pearl (2000), the minimality condition C3 forced the cause to be a single conjunct of the form $X = x$ in every example they considered. They conjectured that it is in fact a consequence of the definition. Hopkins (2002) and Eiter and Lukasiewcz (2001) showed independently that for finite-variable models (the only kind considered in this thesis) condition C3 forces the cause to be a single conjunct of the form $X = x$.

The following example taken from Pearl and Halpern (2000) illustrates the application of their definition of actual causation:

Example (two arsonists) Two arsonists drop lit matches in different parts of a dry forest, and both cause trees to start burning. Consider two scenarios. In the first, "disjunctive" scenario, the lighting of each match is independently sufficient to burn down the whole forest. In the second, "conjunctive" scenario, the forest will burn down only if both matches are lit.

The following causal model describes the "essential structure" of the two scenarios. There are four variables:

- an exogenous variable U which determines, among other things, the motivation and state of mind of the arsonists. For simplicity, assume that $\mathcal{R}(U) = \{u_{0,0}, u_{1,0}, u_{0,1}, u_{1,1}\}$; if $U = u_{i,j}$, then the first arsonist intends to start a fire if and

only if (*iff*) i = 1 and the second arsonist intends to start a fire *iff* j = 1. In both scenarios described here, $U = u_{1,1}$;

- endogenous variables ML_1 and ML_2, each either 0 or 1, where $ML_i = 0$ if arsonist i does not drop the match and $ML_i = 1$ if he does;
- an endogenous variable FB for "forest burns down", with values 0 (the forest does not burn down) and 1 (it does).

The equation for $FB = F_{FB}(U, ML_1, ML_2)$ in the first scenario is: $F_{FB}(u, x, y) = 1$ *iff* x = 1 or y = 1.

In the second scenario, $F_{FB}(u, x, y) = 1$ *iff* $x = 1 \wedge y = 1$. Figure 3.3 is the corresponding causal diagram for both scenarios.

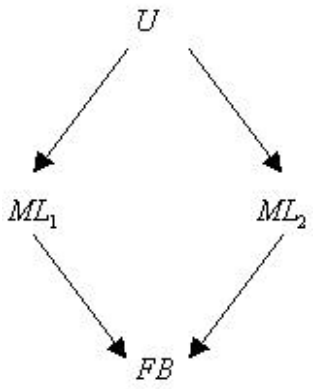

Figure 3.3: Two Arsonists

Let M_1 and M_2 denote the model in the disjunctive and conjunctive scenarios, respectively. Symmetry guarantees that an argument for $ML_1 = 1$ as an actual cause of $FB = 1$ (representing j), in either scenario, gives rise to a symmetrical argument for ML_2. So it is only necessary to consider ML_1: For the (disjunctive) model M_1, let $\vec{Z} = \{ML_1, FB\}$ and therefore $\vec{W} = \mathcal{V} - \vec{Z} = \{ML_2\}$. Even if $ML_1 = 0$, $ML_2 = 1$ gives $FB = 1$. ($U = u_{1,1}$ in both scenarios means that both arsonists intend to start a fire.) Thus $(\vec{Z}, \vec{W})$ = (ML_1, ML_2) must be set to (0,0) to satisfy condition C2(a). When ML_1 is

returned to its original value ($ML_1 = 1$) FB returns to its original value ($FB = 1$), which in this case means that every subset of $\vec{Z}$ returns to its original value even though ML_2 retains its altered value. Thus C2(b) is satisfied. Conditions C1 and C3 are trivially satisfied. Therefore, $ML_1 = 1$ is an actual cause of $FB = 1$ in actual world $(M_1, u_{1,1})$.

For the conjunctive scenario M_2, again let $\vec{Z} = \{ML_1, FB\}$ and $\vec{W} = \{ML_2\}$. In this case, if $ML_1 = 0$ then $FB = 0$ and ML_2 may be left at its original value to satisfy C2(a). This means C2(b) is trivially satisfied, as again are C1 and C3. Therefore, $ML_1 = 1$ is an actual cause of $FB = 1$ in actual world $(M_2, u_{1,1})$.

3.7 The Role of Causal Modelling

Halpern and Pearl (2000) emphasize the critical role that causal modelling plays in actual causal queries. They admit that two closely related structural models for the same system or scenario may give different answers to the same causal query; *A* might be an actual cause of *B* in one model but not in the other. They argue that this is a *feature* of their approach (not a "bug") reflecting that the truth of any claim must be evaluated relative to a particular model of the world:

> It moves the question of actual causality to the right arena—debating which of two (or more) models of the world is a better representation. This, indeed, is the type of debate that goes on in informal (and legal) arguments all the time. (Halpern and Pearl 2000, p. 2)

Among the choices that must be made in modelling some scenario is which variables to treat as endogenous and which to treat as exogenous. Since the values of exogenous variables are assumed to be given, they can be used to represent the background situation, or background assumptions may be implicitly encoded in the structural equations themselves. Halpern and Pearl give an example where the issue is whether a lightning strike or the lighting of a match caused a forest fire. That the wood was dry enough, that there was sufficient oxygen present, and numerous other conditions necessary for a forest fire to occur as a result of either lightning strike or a match are not the focus of interest and are assumed present, part of the background. The modeller chooses to model the dryness by an exogenous variable *D* with values 0 (the wood was

too wet to burn) and 1 (the wood was dry enough), but he does not model the presence of oxygen at all. He models the fire as WB (0: wood is not burning; 1: wood is burning), the lighting of the match as ML (0: the match is not lit; 1: the match is lit), and the occurrence of lightning as L (0: there is no lightning; 1: there is). Then $WB = F_{WB}(\mathcal{U}, \mathcal{V} - \{WB\}) = F_{WB}(D, L, ML)$.

For example, $1 = F_{WB}(1,0,1)$ says that in the context in which the wood is dry enough ($D = 1$), if the match is lit the wood will burn even if there is no lightning. The equation implicitly models the assumption that oxygen is present. Alternatively, the modeller might have included oxygen as an exogenous variable. By not doing so, the modeller does not contemplate contexts in which oxygen is not present; it might, as Pearl and Halpern say, be different if he were modelling scenarios involving fires on Mount Everest.

The choice of endogenous variables reflects not only the choice of which causal mechanisms to represent but which contingencies affecting the model are contemplated. For example, an identical scenario (or part of it) might be modelled by both Figure 3.4(a) and Figure 3.4(b). Figure 3.4(a) does not allow for contingencies affecting C; causal assertions that may depend upon such contingencies might not be accurately represented. Halpern and Pearl consider an example from (Hall 2003) involving railroad tracks in the configuration represented by Figure 3.5(a). There is a switch at X where the track diverges. If the switch is flipped, then the train follows the left-side track. The track converges again at Y and the train eventually ends up at Z whether it follows the left-hand or right-hand track. Hall argues that this example shows the difference between causation and determination: flipping the switch at X cannot cause the train to arrive at Z because it will arrive there in any case; rather, the switch at X merely determines *how* the arrival comes about (by the right-hand or left-hand track).

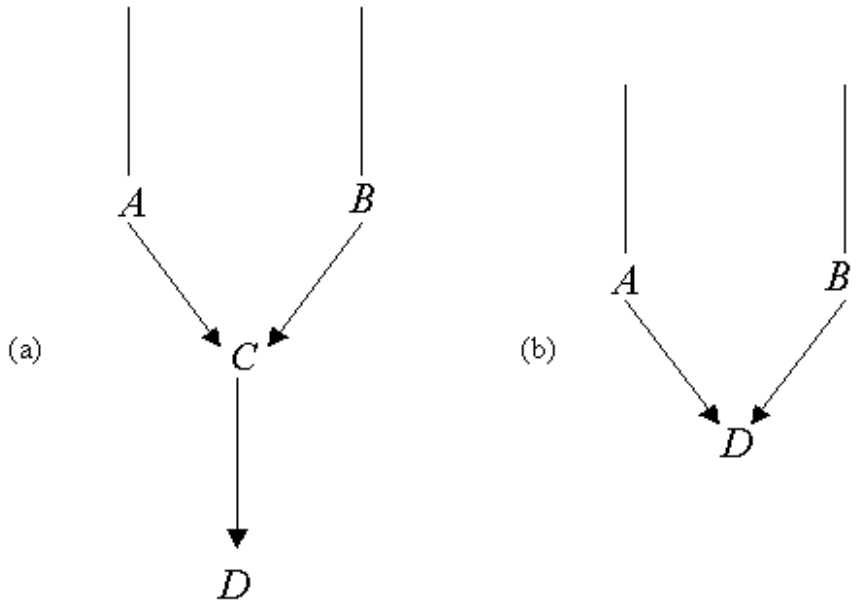

Figure 3.4: Differing Models for the Same Scenario

Halpern and Pearl model this scenario as follows (unless otherwise stated, the context $\vec{u}$ here and subsequently is assumed to be such that it "ensures that the right things happened;" here, that the switch was flipped.):

- *F* for "flipping the switch", with values 0 (the switch is not flipped) and 1 (it is);
- *T* for "track", with values 0 (the train follows the left track) and 1 (the train follows the right track); and
- *A* for "arrival", with values 0 (the train does not arrive at the point of re-convergence) and 1 (it does).

(The corresponding causal diagram is Figure 3.5(b).)

It is easy to see that the value of *A* does not depend on the value of *F* since $A=1$ whether $T=0$ or $T=1$. In other words, condition C2(a) fails and flipping the switch does not actually cause the train to arrive. This is a valid model of Hall's scenario, which does not contemplate contingencies (interventions) preventing the train's arrival once it reaches the switch: the trains *will* arrive since the tracks converge.

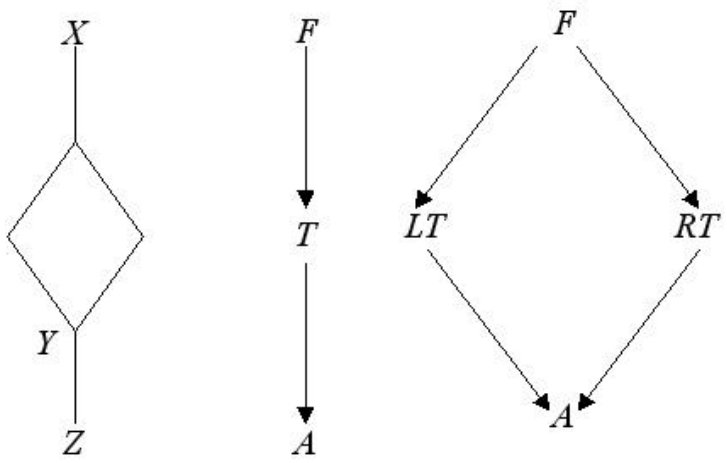

Figure 3.5: Three Train Models

Halpern and Pearl propose an alternative model, replacing the variable *T* with the following two variables:

- *LT* for "left track", with values 1 (the train follows the left track) and 0 (it does not); and
- *RT* for "right track", with values 1 (the train follows the right track) and 0 (it does not).

(The corresponding causal diagram is Figure 3.5(c).) Letting $\vec{Z} = \{F, LT, A\}$ and $\vec{W} = \{RT\}$ it is easy to see that, in this model, $F = 1$ is an actual cause of $A = 1$. This model depicts the tracks as separate mechanisms and therefore contemplates (or advertises) contingencies interfering with the tracks (e.g., a landslide). If the switch is not flipped ($F = 0$) and the left-track mechanism is perturbed ($do(RT = 0)$), then the train will not arrive ($A = 0$). Because of the possibility of such contingencies, flipping the switch ($F = 1$) is a cause of the train's arrival ($A = 1$). According to Halpern and Pearl (2000, p. 24),

> Causal models earn their value in abnormal circumstances, created by structural contingencies, such as the possibility of a malfunctioning track. It is this possibility that should enter our mind whenever we decide to designate each track as a separate mechanism (i.e., equation) in the model

and, keeping this contingency in mind, it should not be too odd to name the switch position a cause of the train arrival (or non-arrival).

3.8 The Halpern-Pearl Definition and Formalizing the NESS Test

As suggested in the introduction to this chapter, there is an almost straightforward interpretation for the NESS test in Pearl's structural language. Recall, according to Wright's description of the application of the NESS test (Wright 1988, p. 1042; see Section 2.3.6), that the purpose of the NESS test is to "determine which causal generalizations have been instantiated in the actual world by the conditions that occurred on the particular occasion." The relative part of the actual world is readily identified with a causal world, a causal model with a particular context. The causal model itself is a complex of causal generalizations, incompletely specified causal "laws" (see Section 2.3.5) as determined by the choice of model variables (which variables to include or exclude, which included variables to treat as endogenous or exogenous). The causal generalizations that have been instantiated on a particular occasion naturally suggest the sets *Z* of the Halpern-Pearl definition. The sets *Z*, active causal processes or "sustaining" sets of variables, appear to satisfy the concept of "actual sufficiency" Wright is striving for in defining the NESS test and inadequately (according to Pearl; see Section 2.3.6) realizing using the language of logical necessity and sufficiency.

However, Hopkins and Pearl (2003) have shown that even locally, between a variable (the effect) and its parents (direct causes), the Halpern-Pearl definition does not require that an active causal process (a set *Z* satisfying condition C2 of the definition) be actually sufficient; the ability of an active causal process to produce the effect in question, as tested by condition C2(b), is (unintentionally) allowed to depend on *non-actual* conditions.

To see this, consider a causal model *M* with context $\mathcal{U} = \vec{u}$. For simplicity, assume that all $U \in \mathcal{U}$ are trivial in the structural equation F_X for X (i.e., U_X is empty; see Section 3.4) so that $F_X : Dom(PA_X) \rightarrow Dom(X)$. Now consider an assignment $\overrightarrow{PA}_X \leftarrow \overrightarrow{pa}_X$ (recall that the vector notation is used in this context to represent an ordered assignment—see Section 3.4) such that $X = x$. If $\overrightarrow{PA}_X = \{V_1, \ldots, V_n\}$ and

$\overrightarrow{pa}_X = \{v_1,\ldots,v_n\}$ then the logical sentence consisting of a conjunction of the literals $V_i = v_i$ (i.e., $V_1 = v_1 \wedge \ldots \wedge V_n = v_n$) implies $X = x$. If such a sentence is formed for each value assignment $\overrightarrow{PA}_X \leftarrow \overrightarrow{pa}_X$ such that $X = x$ and a new sentence, denoted $\Delta(X = x)$, is formed as a disjunction of all such sentences (so that the resulting logical sentence is in disjunctive normal form), then $X = x$ *iff* $\Delta(X = x)$. To illustrate, Hopkins and Pearl give the following example.

Example (firing squad) There is a firing squad consisting of two shooters *B* and *C*, one of whom, *B*, is too lazy to load his own gun. Shooter *C* loads and shoots his own gun while shooter *B* has *A* load his gun for him. The prisoner *D* will die (*D* = 1) if, and only if, either *A* loads *B*'s gun (*A* = 1) and *B* shoots (*B* = 1) or *C* loads his gun and shoots (*C* = 1); that is, $D = (A \wedge B) \vee C$.

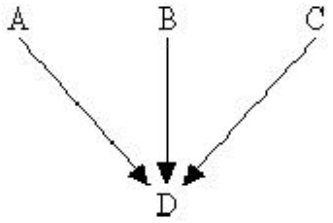

Figure 3.6: Firing Squad

For this example, if $(M, \vec{u}) \models (D = 1)$ then

$$\Delta(D=1) = (A=1 \wedge B=1 \wedge C=1) \vee (A=1 \wedge B=1 \wedge C=0) \vee (A=0 \wedge B=1 \wedge C=1) \vee (A=1 \wedge B=0 \wedge C=1) \vee (A=0 \wedge B=0 \wedge C=1)$$

A term (conjunction of literals) that entails a sentence *S* is an *implicant* of *S*; an implicant that does not entail any other implicant is a *prime implicant*. The *prime implicant form* of a sentence is a disjunction of all its prime implicants and is unique. The prime implicant form of $\Delta(D = 1)$ is $\Delta(D = 1) = (A = 1 \wedge B = 1) \vee (C = 1)$.

With these preliminaries, Hopkins and Pearl (2003) prove the following theorem:

Theorem (prime implicant) In a causal model *M* with context $\mathcal{U} = \vec{u}$, let $X, Y \in \mathcal{V}$ with $X \in PA_Y$. If $(M, \vec{u}) \models (X = x \wedge Y = y)$ and the literal $X = x$ occurs in

any prime implicant of $\Delta(Y = y)$ then the Halpern-Pearl definition of actual causation will classify $X = x$ as an actual cause of $Y = y$.

Note that the prime implicant theorem does *not* require that any other literals (if they exist) in any of the prime implicants of $\Delta(Y = y)$ to which $X = x$ belongs be satisfied (true) in $(M, \vec{u})$. For example, assume the context $\vec{u}$ in the firing squad example is such that C shoots and A loads B's gun, but B does not shoot. Since $(M, \vec{u}) \models (A = 1 \wedge D = 1)$ and $A = 1$ occurs in the prime implicant $(A = 1 \wedge B = 1)$ for $\Delta(D = 1)$, according to the prime implicant theorem, the Halpern-Pearl theorem should (counter-intuitively) classify A's loading of B's gun as a cause of D's death though B does not shoot. Indeed, taking $\vec{Z} = (A, D)$ and $\vec{W} = (B, C)$ and setting $\vec{W} = \vec{w} = (1,0)$) satisfies conditions C2(a) and (b) of the definition.

Hopkins and Pearl (2003) point out the similarity of the prime implicant form of a sentence with Mackie's INUS condition (see Section 2.3.4):

> For instance, A loading B's gun is a necessary part of a sufficient condition to ensure the prisoner's death. In terms of the prime implicant logical form, sufficient conditions map to implicants. For instance, $A = 1 \wedge B = 1$ is a sufficient condition for $D = 1$. Furthermore, since $A = 1 \wedge B = 1$ is a *prime* implicate[13] (hence no subset of its conjuncts is an implicate), we observe that both $A = 1$ and $B = 1$ are necessary parts of this sufficient condition. Hence any atomic expression that appears in a prime implicate satisfies the INUS condition.

Accepting the mapping of sufficient conditions to implicants, then Mackie's analysis (see Section 2.3.5) requires that for $A = 1$ to be a cause of $D = 1$, not only must $A = 1$ occur as an atomic proposition (or literal) in some prime implicant for $D = 1$ (i.e., be an INUS condition for $D = 1$) but also that every other atomic proposition in that implicant be satisfied. Recall also that this part of Mackie's analysis is consistent with the NESS test. It follows then that the Halpern-Pearl definition is less restrictive than both Mackie's INUS analysis and the Wright's NESS test. As Hopkins and Pearl say, their prime implicant theorem exposes that the Halpern-Pearl definition is over permissive. It is at least too permissive to formally capture the meaning of the NESS test.

[13] Hopkins and Pearl are apparently using "implicate" synonymously with "implicant" here.

Having shown that the Halpern-Pearl definition is too permissive, Hopkins and Pearl (2003) go on to question the general validity of the "counterfactual strategy"—that "event *C* causes event *E iff* for some *appropriate G*, *E* is counterfactually dependent on *C* when we hold *G* fixed." They also question whether it is possible at all, given what they believe to be representational limitations of the structural language, to give a satisfactory definition of actual causation within the structural model framework. In response, the next chapter of this thesis develops a new structural definition of actual causation, inspired by the manner in which the Halpern-Pearl definition fails to capture the NESS test (as shown by Hopkins and Pearl), and with it argues that the Hopkins and Pearl critique of the counterfactual strategy and the so-called limitations of the structural language are not conclusive.

4. A New Structural Definition of Actual Causation

This chapter develops an alternative structural definition of actual causation and defends the validity of the counterfactual strategy, and the structural mode approach in general, from criticisms in Hopkins and Pearl (2003). The definition of actual causation developed in this chapter continues the basic counterfactual strategy while avoiding problems with this approach identified by Hopkins and Pearl. This new definition attempts to formalize Wright's NESS test in the structural language and differs from the Halpern-Pearl (2000) approach by syntactically encoding causal information in causal models interpretable as describing sufficient conditions or sets of sufficient conditions for some effect.

This chapter is organized as follows. Section 4.1 argues that the Halpern-Pearl approach to defining actual causality ignores important causal information encoded in structural equations. Section 4.2 attempts to elicit what that causal information is with the help of the concept of "coefficient invariance" developed by Hausman and Woodward (1999). Section 4.3 introduces a concept derived from Hausman and Woodward's explanation of coefficient invariance that is then used to develop a new definition of actual causation in Section 4.4. Section 4.5 considers objections raised by Hopkins and Pearl (2003) to the general counterfactual strategy for defining actual causation in the language of structural models. Finally, Section 4.6 considers what Hopkins and Pearl (2003) describe as "ontological concerns" that they suggest call into question the suitability of the causal model framework for defining actual causality.

4.1 Recalling Lost Structure

Recall (Section 3.3) that a structural equation $x = f_x(pa_x, e_x)$ for x represents a causal mechanism determining the value of x where pa_x represents modelled direct causes (parents in the corresponding graphical representation) and e_x represents non-modelled

factors directly affecting x. For a given set of values for pa_x and e_x, the structural equation for x defines an equilibrium state (Pearl 2000). What makes the equation "structural" is that under an intervention that changes the value of some $Y \in pa_x$ from the (pre-intervention) value $Y = y$ to $Y = y'$ ($y \neq y'$), in the resulting new equilibrium state the relationship described by f_x between X, pa_x and e_x continues to hold. Hausman and Woodward (1999) call this invariance property *level invariance*. A system of equations that admits a structural interpretation (as a causal model) must satisfy level invariance and what Hausman and Woodward call *modularity* (cf. Pearl's *autonomy*; see Section 3.2):

> It says that each structural equation in a system of structural equations that correctly captures the causal relation among a set of variables is invariant under interventions that disrupt other equations in the system by setting the values of their dependent variables....

The effect of fixing the value of an endogenous variable X to x is to replace the equation for X with the constant function $X = x$. Graphically, this corresponds to breaking all arrows directed into X in a graphical representation of a system of equations (model) including the equation for X. Modularity means that this does not break arrows directed from X to other endogenous variables in the model.

A system of equations satisfying level invariance and modularity encodes an *interventional function* (Hopkins and Pearl 2003); that is, for a particular causal world (i.e., causal model with a specified context, a setting for the exogenous variables) it is possible to determine the value of any endogenous variable given that some other endogenous variables have their value fixed at some non-actual values (i.e., an intervention). It is the interventional function that gives causal models their ability to answer counterfactual queries (see Section 1.5) and makes the structural language attractive for formalizing a counterfactual definition of actual causation, as in the Halpern-Pearl definition.

Hopkins and Pearl (2003) suggest that it is only in the choice of a causal model's endogenous variables and the corresponding interventional function that a causal world "essentially" encodes causal information. Pearl (2000) points out, however, the

importance of the structural information conveyed by structural equations. Consider the desert traveller example (see Section 2.3.5), where a traveller has two enemies, one who poisons (p = 1) the traveller's water canteen and the other, unaware of the poisoning, shoots and empties the traveller's canteen (x = 1) as a result of which the traveller dies. Pearl (2000, p. 312) considers the structural equations $y = x \vee x'p$ (x' is equivalent to $\neg x$) and the *logically* equivalent $y = x \vee p$ and states,

> Here we see in vivid symbols the role played by structural information. Although it is true that $x \vee x'p$ is logically equivalent to $x \vee p$, the two are not structurally equivalent; $x \vee p$ is completely symmetric relative to exchanging x and p, whereas $x \vee x'p$ tells us that, when x is true, p has no effect whatsoever—not only on y, but also on any of the intermediate conditions that could potentially affect y. It is this asymmetry that makes us proclaim x and not p to be the cause of death.

Whatever this structural information is, it plays no part in Halpern and Pearl's (2000) explication of their definition of actual causation. Consider their analysis of the following example from Hall (2003).

Example (rock-throwing) Suzy and Billy accurately throw rocks at the same bottle and with sufficient force to shatter the bottle. Suzy's throw arrives first and shatters the bottle before Billy's throw can arrive and shatter the bottle.

Intuitively, Suzy's throw causes the bottle to shatter and not Billy's. Halpern and Pearl first consider a "coarse" model with three propositional variables *ST* (Suzy Throws), *BT* (Billy Throws), and *BS* (Bottle Shatters) and single structural equation $BS = ST \vee BT$. Halpern and Pearl (2000, p. 14) say,

> In this simple causal network, *BT* and *BS* play absolutely symmetric roles, with $BS = ST \vee BT$, and there is nothing to distinguish one from the other. Not surprisingly, both Billy's throw and Suzy's throw are classified [by the Halpern-Pearl definition] as causes of the bottle shattering.

It is not surprising since, by analogy with the desert traveller example, it is the asymmetry between the roles played by *ST* and *BT* in the actual scenario that "makes us proclaim" *ST* and not *BT* to be the cause of the bottle shattering. $BS = ST \vee BT$ is a truth-conditional equation for *BS*, distinct from the structural equation

$BS = ST \vee (\neg ST \wedge BT)$. Indeed, Halpern and Pearl in applying their definition to example scenarios frequently do not even bother to explicitly provide structural equations for their models, instead implying a truth-conditional equation based intuitively on the provided model variables.

Suppose that Billy's throw was actually earlier than Suzy's and would have shattered the bottle if not for the presence of a translucent net ($N = 1$) in the path of Billy's throw. It is still the case, logically, that $BS = ST \vee BT$ but it is harder to confuse that logical equation with the structurally distinct $BS = (ST \vee N) \vee (BT \wedge \neg N)$. Yet, with respect to the causal structure of the scenario, ST plays an analogous role to that of N in the modified scenario.

Perhaps the reason Halpern and Pearl are insensitive to the structural information conveyed by structural equations is that their definition is insensitive to it. For example, if the equation for BS in the Halpern and Pearl model for the rock-throwing example is replaced by the structurally distinct $BS = ST \vee (\neg ST \wedge BT)$ then the Halpern-Pearl definition will still classify $BT = 1$ as a cause of $BS = 1$ even when $ST = 1$ (for condition C2 of the Halpern-Pearl definition take $\vec{Z} = \{BT, BS\}$, $\vec{W} = \{ST\}$, and set $ST = 0$). Generally, the insensitivity of the Halpern-Pearl definition of actual causation to the structural information that the causal relation between two variables depends on the value of one or more other variables is a corollary of the Hopkins and Pearl prime implicant theorem (see Section 3.8).

Thus, while it may be true that the choice of endogenous variables and the corresponding interventional function (the structural contingencies) comprise the information encoded in causal worlds *relevant* to the Halpern-Pearl definition's determination of actual causation, it is not all the information that is encoded. The next section will consider what meaning can be attached to the structural information that the causal relation between an independent variable and dependent variable in a structural equation depends (or does not depend) on one or more other independent variables in the equation (e.g., the difference between $BS = ST \vee BT$ and $BS = ST \vee (\neg ST \wedge BT)$) and what, if any, role that information plays in the determination of actual causation.

4.2 Coefficient Invariance

A key to understanding the meaning of the structural information encoded in structural equations is the realization that a single structural equation can describe more than one distinguishable mechanism. As Hausman and Woodward (1999, p. 548 n. 15) state,

> Although we speak of the mechanism an equation captures, we do not interpret modularity[14] as ruling out the possibility that a single structural equation might describe the operation of more than one mechanism. In the case of additive relations a single structural equation may express several mechanisms that are distinct from one another and which can be separately disrupted.

If in a system of structural equations with additive terms each *term* in each equation represents a distinct causal mechanism ("and thus acts independently of the mechanisms registered by other terms") Hausman and Woodward (1999, p. 547) say that the system exhibits *coefficient*[15] *invariance*:

> Coefficient invariance is a restrictive condition: it will be violated whenever additivity is, or when the causal relationship between two variables depends on the level of a third variable. If one thinks of individual causes of some effect Y as conjuncts in a minimal sufficient condition for Y (or the quantitative analogue thereof)—that is, as 'conjunctive causes'—then the relationship between an effect and its individual causes will not satisfy coefficient invariance. Removing the arrow between an individual cause X and one of its effects Y will not leave the coefficients relating Y to its other causes unaffected. Coefficient invariance can be expected to hold only when the vertices in a graph that represent causes of Y in fact represent (components of) *separate* minimal sufficient conditions—i.e., 'disjunctive causes.'

There is something of a contradiction in the way Hausman and Woodward explain coefficient invariance; at least, as a means of identifying distinguishable (if not distinct in the sense of shared variables) mechanisms within a single structural equation, it is too strict. To illustrate this, recall Hopkins and Pearl's firing squad example (see

[14] The sense in which Hausman and Woodward use "modularity" here includes level invariance (see Hausman and Woodward 1999, p. 545).

[15] "Coefficient" invariant because it requires that "individual coefficients be *separately* invariant under interventions that change the value of other coefficients" (Hausman and Woodward 1999, p. 547). While it is not critical for this discussion, for a discussion of what it means for an intervention to change the value of a coefficient see (Hausman and Woodward 1999, Section 5).

Section 3.8) with structural equation $D = (A \wedge B) \vee C$. $(A \wedge B)$ and C represent distinct mechanisms (minimal sufficient conditions) for the prisoners death, *D*. However, for the term $A \wedge B$ the existence of a causal relation between B ($B = 1$) and D ($D = 1$) depends on the value of *A*. Hausman and Woodward's explanation would require the minimal sufficient conditions that represent separate individual causes consist of a single variable; that is, the individual terms in the structural equations would contain a single variable and the structural equations would be sums of these terms. This is confirmed by Hausman and Woodward when they say (1999, p. 547), "When one has additive relations, one can interpret individual edges as representing separate mechanisms."

Even requiring that the mechanisms represented by individual terms act "independently of the mechanisms registered by other terms" is too restrictive. In the case of the rock-throwing example (see Section 3.8), where the structural equation is $BS = ST \vee (\neg ST \wedge BT)$, the terms describe separate minimal sufficient conditions for the bottle shattering ($BS = 1$). This is even clearer with another example with the same causal structure, Pearl's (2000, p. 311) two-switches example:

> Consider an electrical circuit consisting of a light bulb and two switches, as shown in…[Figure 4.1]. From the user's viewpoint, the light responds symmetrically to the two switches; either switch is sufficient to turn the light on. Internally, however, when switch 1 is on it not only activates the light but also disconnects switch 2 from the circuit, rendering it inoperative.

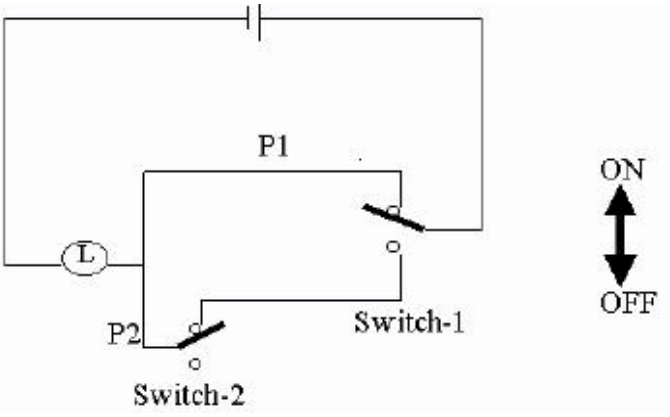

Figure 4.1: Pearl's (2000) Two Switches scenario.

With (propositional) variables $S1$ ("Switch 1"), $S2$ ("Switch 2"), and L ("Light") the structural equation is $L = S1 \vee (\neg S1 \vee S2)$ (i.e., the Light is "ON" when Switch 1 is "ON" or Switch 1 is "OFF" and Switch 2 is "ON"). It seems clear that the disjuncts $S1$ and $\neg S1 \vee S2$ in the equation for L represent distinguishable mechanisms. When Switch 1 is "ON," the mechanism by which L is "ON" involves power flowing through path $P1$ (see Figure 4.1); it does not matter whether the path $P2$ from Switch 1 to L is connected (i.e., Switch 2 is "ON") or not. On the other hand, if Switch 1 is "OFF" then the mechanism by which L is "ON" involves power flowing through path $P2$ only and path $P1$ could be cut, or removed, without effect. Though the mechanisms cannot act independently, in the sense that only one can be active at a time, they are distinguishable. Distinct mechanisms can share components (variables).

The requirements of separate disruption of individual terms in a structural equation and additivity of terms (i.e., the causal relationship between a variable in a term and the dependent effect variable cannot depend on the level of another variable in the same term) are more restrictive than necessary to have distinct terms in additive structural equations represent distinct mechanisms (separate minimal sufficient conditions). Suppose instead that the requirement that "the vertices in a graph that represent causes of Y in fact represent (components of) *separate* minimal sufficient conditions—i.e., 'disjunctive causes'" is interpreted literally to mean that for a set of additive structural equations, expressed in sum of products form,[16] each term in an equation represents a separate mechanism, a separate set of minimal sufficient conditions ("or the quantitative analogue thereof"). Call this property *term modularity*. The next section uses this property to originate and develop a concept of *relative* coefficient invariance as a criterion for deciding what variables should be held fixed and what variables may be altered in testing for counterfactual dependence between a effect variable and one of its (putative) causal variables in the model. This, in turn, will allow for a new structural definition of actual causation that avoids the problem identified by Hopkins and Pearl (see Section 3.8) and that formalizes the NESS test in the structural language for a scenario modelled by a causal world.

[16] Recall that a disjunction ("$\vee$") is a Boolean sum and a conjunction ("$\wedge$") is a Boolean product.

4.3 coin $_{\vec{u}}$ (Z;X|Y)

For the parents of a variable, the "term modularity" criterion encodes the relationship of being components of distinct component causal mechanisms (or elements of minimal sufficient conditions for the variable). For distinct variables *X*, *Y*, and *Z*, where *X* and *Y* occur as independent variables in the structural equation (parents) for Z in a causal world $(M,\vec{u})$, X *is coefficient invariant to* Y *for term* T if (1) $(M,\vec{u}) \models \neg(T = 0)$ (i.e., the term is satisfied, or non-zero in quantitative contexts, in $(M,\vec{u})$), (2) *X* occurs as a literal in *T*, and (3) *Y* is not a variable in *T* (symbolically, $coin_{\vec{u}}^{T}(Z; X \mid Y)$; note that *X* is a literal of the form $X = x$ where $(M,\vec{u}) \models (X = x)$ while *Y* is a variable). When $coin_{\vec{u}}^{T}(Z; X \mid Y)$ the causal relation between *X* and *Z* does not depend on the value of *Y* in context $\vec{u}$. The reason for developing this definition is to avoid satisfying actually unsatisfied terms (minimal sufficient sets) when changing the values of variables not in the causal process being tested, as happens with the Halpern-Pearl definition (see Section 3.8). This is accomplished by requiring that between a variable *X* and its parents $(Y_1,\ldots,Y_n)$, in testing whether $Y_i = y_i$ is an actual cause of $X = x$, $Y_i = y_i$ should belong to a satisfied term *T* and only parent variables Y_j that Y_i is coefficient invariant to for *T* in the equation for *X* $(coin_{\vec{u}}^{T}(X; Y_i \mid Y_j))$ should be allowed to have their values altered.

Returning to Hopkins and Pearl's firing squad example (see Section 3.8), with structural equation $D = (A \wedge B) \vee C$ and context such that $A = C = 1$ and $B = 0$, $(M,\vec{u}) \models \neg(A \wedge B)$ and it is not the case that *A* is coefficient invariant to *B* for $T = (A \wedge B)$ in the equation for *D* $(\neg coin_{\vec{u}}^{(A \wedge B)}(D; A \mid B))$ because *T* is not satisfied. Therefore, in the context such that $A = C = 1$ and $B = 0$, to test whether $D = 1$ is counterfactually dependent on $A = 1$, the value of *B* may not be altered. Since $(A \wedge B)$ is the only term in the equation for *D* in which *A* (i.e., $A = 1$) occurs, contrary to the Halpern-Pearl approach, it is not possible to modify the model so that *D* is counterfactually dependent on *A* in a scenario where $B = 0$.

Globally in a causal model with context $\mathcal{U} = \vec{u}$, when distinct variables Y_1 and Y_2 occur as common parents of distinct variables X_1 and X_2 it can happen that there exists a term T_{X_1} in the equation for X_1 such that $coin_{\vec{u}}^{T_{X_1}}(X_1; Y_1 | Y_2)$ but any satisfied term T_{X_2} in the equation for X_2 that includes Y_1 also includes Y_2 ($\neg coin_{\vec{u}}^{T_{X_2}}(X_2; Y_1 | Y_2)$); that is, Y_1 is coefficient invariant to Y_2 for some term in the equation for X_1 but not in the equation for X_2. In that case, if Y_1 is part of the "active causal process" being tested, before allowing the value of Y_2 to be altered, it is necessary to interfere directly in the equation for X_2 by substituting a constant y_2 for Y_2 in the equation for X_2 where $Y_2 = y_2$ in the unaltered model (i.e., fix Y_2 at its actual value, $(M, \vec{u}) \models (Y_2 = y_2)$). This avoids the possibility of the counterfactual or original values of Y_1 interacting with non-actual values to satisfy non-actually satisfied minimal sufficient sets for some variable, the problem that plagues the Halpern-Pearl definition (see Sections 3.8 and 4.5). This process must be repeated for all X_i where for all satisfied term T_{X_i} including $\neg coin_{\vec{u}}^{T_{X_i}}(X_i; Y_1 | Y_2)$. Only then should altering the value of Y_2 be allowed.

4.4 A New Structural Definition of Actual Causation

Before showing how these concepts can be applied to a new definition of actual causation two further definitions are required:

A *causal route* $\vec{R} = \langle C, D_1, \ldots, D_n, E \rangle$ between two variables C and E in $\mathcal{V}$ is an ordered sequence of variables such that each variable in the sequence is in $\mathcal{V}$ and a parent of its successor in the sequence.

The following (original) definition deals with the issue that Y_1 may be coefficient invariant to Y_2 for some term in the equation for X_1 but not in the equation for X_2. For a causal mode M with route $\vec{R} = \langle C, D_1, \ldots, D_n, E \rangle$ and a sequence of terms

$\vec{T} = \langle T_{D_1}, \ldots, T_{D_n}, T_E \rangle$, where T_X is a satisfied term in the equation for X, the *submodel relative to* $\vec{R}$ *and* $\vec{T}$ *in context* $\vec{u}$ (denoted $M^{\vec{T}}_{[\vec{R},\vec{u}]}$) is derived from $(M, \vec{u})$ as follows: for distinct $X, Y, W \in \mathcal{V}$ with $X \in \vec{R} - E$, $Y \notin \vec{R}$, and $W \neq C$, if $\neg coin^{T_W}_{\vec{u}}(W; X \mid Y)$ replace the function F_W for W (see Section 3.4) by the function that results when V is replaced with a constant v where $(M, \vec{u}) \models (V = v)$.

Definition (*actual cause*; *new version*) $C = c$ is an actual cause of $E = e$ in $(M, \vec{u})$ if the following conditions hold:

AC1. $(M, \vec{u}) \models (C = c \wedge E = e)$

AC2. There exists a route $\vec{R} = \langle C, D_1, \ldots, D_n, E \rangle$ in M, a sequence of satisfied terms $\vec{T} = \langle T_{D_1}, \ldots, T_{D_n}, T_E \rangle$, and a setting $\vec{w}$ for $\vec{W} = \mathcal{V} - \vec{R}$ and a setting $c' \neq c$ for C such that:

(a) $(M^{\vec{T}}_{[\vec{R},\vec{u}]}, \vec{u}) \models [C \leftarrow c', \vec{W} \leftarrow \vec{w}] \neg (E = e)$, and

(b) $(M^{\vec{T}}_{[\vec{R},\vec{u}]}, \vec{u}) \models [C \leftarrow c, \vec{W} \leftarrow \vec{w}](E = e)$.

Because there are no causal interaction effects between variables in $\vec{R}$ and $\vec{W}$ in $M^{\vec{T}}_{[\vec{R},\vec{u}]}$, by the construction of $M^{\vec{T}}_{[\vec{R},\vec{u}]}$ (variables in $\vec{R}$ are coefficient invariant to all variables in $\vec{W}$ by definition of $M^{\vec{T}}_{[\vec{R},\vec{u}]}$), the setting $\vec{W} \leftarrow \vec{w}$ cannot "contaminate" the test of counterfactual dependence in AC2 in the sense of satisfying a non-actually satisfied minimal sufficient set of conditions.

In practice, it rarely happens that a literal X occurs in more than one satisfied term in a structural equation; a non-quantitative equation having more than one satisfied term with distinct literals only occurs itself in cases of duplicative causation (see Section 1.1). To avoid the cumbersome and somewhat confusing terminology, subsequently, unless the context requires otherwise (as in the analysis of the pollution cases in Section

5.3), the choice of the sequence $\vec{T}$ will be left as implied by the analysis of the scenario and the superscript $\vec{T}$ left out of the notations $coin_{\vec{u}}^{T}(Z; X \mid Y)$ and $M_{[\vec{R},\vec{u}]}^{\vec{T}}$.

As an example of how this definition is applied, consider the rock-throwing example of Section 4.1 with the single structural equation $BS = ST \vee (\neg ST \wedge BT)$[17]. To show that $ST = 1$ is a cause of $BS = 1$, let $\vec{R} = \langle ST, BS \rangle$. Since $(M, \vec{u}) \models (ST = 1)$, the term consisting of the single literal ST is satisfied in $(M, \vec{u})$. Therefore $coin_{\vec{u}}(BS; ST \mid BT)$ and the model's only structural equation is unchanged in $M_{[\langle ST, BS \rangle, \vec{u}]}$. Since $(M, \vec{u}) \models (ST = 1 \wedge BS = 1)$, condition AC1 is satisfied. For condition AC2, $\vec{W} = \mathcal{V} - \vec{R} = \{BT\}$. Setting $BT \leftarrow 0$ ($\vec{W} \leftarrow \vec{w}$) and $ST \leftarrow 0$ ($C \leftarrow c'$) satisfies condition AC2(a) (i.e., if $ST = BT = 0$ then $BS = 0$). Keeping $BT = 0$ and setting ST back to its actual value, $ST \leftarrow 1$, results in $BS = 1$ satisfying condition AC2(b) . Thus the definition classifies $ST = 1$ as a cause of $BS = 1$.

To show that $BT = 1$ is *not* a cause of $BS = 1$ in $(M, \vec{u})$, take $\vec{R} = \langle BT, BS \rangle$, the only route from *BT* to *BS*. Note that the only term containing *BT* (i.e., $BT = 1$), $\neg ST \wedge BT$, is not satisfied in $(M, \vec{u})$ as $(M, \vec{u}) \models \neg(\neg ST \wedge BT)$. Therefore, $\neg coin_{\vec{u}}(BS; BT \mid ST)$ and $ST \leftarrow 1$ in the equation for *BS* in $M_{[\langle BT, BS \rangle, \vec{u}]}$ where the equation for *BS* becomes $BS = 1 \vee (0 \wedge BT) = 1$. Clearly it is not possible for condition AC2 to be satisfied; the definition will not classify $BT = 1$ as a cause of $BS = 1$.

(Note that, strictly speaking, the equation for condition for *BS* would be of the form $BS = (ST \wedge \neg u'_{ST}) \vee (\neg ST \wedge BT \wedge \neg u'_{(\neg ST \wedge BT)}) \vee u_{BS}$ where u'_X stands for inhibiting abnormalities; for example, u'_{ST} stands for inhibiting abnormalities that would prevent the bottle shattering from Suzy's throwing the rock. u_{BS} stands for triggering abnormalities that might cause the bottle to shatter even if neither Suzy nor Billy throw their rocks (see Pearl 2000, p. 29). Thus in $M_{[\langle BT, BS \rangle, \vec{u}]}$ the equation for *BS*

[17] Recall Suzy and Billy throw rocks (*ST* and *BT*) at a bottle and Suzy's arrives first to shatter the bottle (*BT*) before Billy's can arrive to do the same.

would becomes $BS = u'_{ST} \vee u_{BS}$. Generally, unless otherwise stated, the assumption is that these abnormalities are not present and are left out of the equations.)

4.5 Validity of the "Counterfactual Strategy"

The counterfactual strategy—"Event *C* causes event *E* iff for some *appropriate G*, *E* is counterfactually dependent on *C* when we hold *G* fixed"—is at least as old as Mackie's counterfactual account of causation as that which makes a difference in relation to some background or causal field (Mackie 1974). The strategy is pursued famously by Lewis (1979) with his account of what must be held fixed in applying his similarity metric for possible worlds (see Section 2.3.3), and is pursued as recently as Yablo's (2003) account of effects being *de facto* dependent on their causes *modulo* some fixed infrastructure. According to Halpern and Pearl, the key element of the counterfactual strategy is the identification of which *G* are appropriate to hold fixed: "Intuitively, we would like to screen out the other causes of E, such that the only causal mechanism responsible for E is C. In the case of the Halpern-Pearl definition (see Section 3.6), condition C2 defines *G* as a particular setting, relative to a causal model $(M, \vec{u})$, of a group of endogenous variables ($\vec{W}$) not in the "active causal process" ($\vec{Z}$) that together with the active causal process partition the endogenous variables ($\mathcal{V} = \vec{W} \cup \vec{Z}$) so that C2(a) and C2(b) are satisfied.

That condition C2 of the Halpern-Pearl definition is unintuitive does not need argument (and Hopkins and Pearl do not provide one; though by comparison with Pearl's (2000) original "causal beam" definition, the Halpern-Pearl definition is positively crystalline). In arguing that the restrictions on possible *G*,—embodied in condition C2(b)—are too permissive, Hopkins and Pearl (2003) rely on two example scenarios for which they claim the Halpern-Pearl definition gives counterintuitive causal answers. In what follows, it is argued that the first ("loanshark") example considered by Hopkins and Pearl (Section 4.5.1) exploits the same flaw in the definition identified in their prime-implicant theorem that the new definition of actual causation ("new definition") eliminates. Additionally, in Section 4.5.2 it is argued that Hopkins and Pearl's use of the second ("bomb") example illustrates how their approach to

undermining the counterfactual strategy ("the problem illustrated by this example is simply a representative of any number of situations where it is possible to choose an inappropriate G to keep fixed") fails to distinguish between the distinct issues of inappropriate G's and inappropriate causal models.

4.5.1 Condition C2(a) or C2(b)¾ which is too permissive?

Recall that Halpern and Pearl (2000) intended that the variables in the set $\vec{Z}$ of condition C2 of their structural model definition of actual causation should mediate between the (putative) cause $X = x$ and the effect in question, j, and that the set $\vec{Z}$ should be thought of as describing the "active causal process" from X to j (see Section 3.6). The choice of the set $\vec{W}$ and the setting $\vec{W} = \vec{w}'$ is intended to isolate this causal process in testing the causal influence of X by shutting off other potential causal processes for j; that is, for some setting $x \neq x'$, when $X = x'$ and $\vec{W} = \vec{w}'$, j is false ($\neg j$, condition C2(a)). Condition C2(b) is then intended to show that the causal process from $X = x$ to j is "active": other causal processes (if any) are shut off ($\vec{W} = \vec{w}'$) and the presumed causal process from $X = x$, mediated by $\vec{Z}$, is shut off ($X = x'$) so that when the latter process is restarted ($X = x$), should j become true again it can be attributed to the causal influence of $X = x$.

Halpern and Pearl (2000) recognize, however, that if it is to be ensured that the original causal process mediated by $\vec{Z}$ in the unaltered model is the same process that produces j in the altered model (the submodel $M_{\vec{W} \leftarrow \vec{w}'}$) then any influences on $\vec{Z}$ resulting from the interventions in the model described by the setting $\vec{W} \leftarrow \vec{w}'$ must be screened. Therefore the requirement of condition C2(b) that not only should returning $\vec{X}$ to its original value $X = x$ make j true again but also restoring the original values of any subset of variables in $\vec{Z}$ should leave j true. However, condition C2(b) is not only unintuitive, it is insufficient. Halpern and Pearl apparently failed to recognize that condition C2(a) does not ensure that $\neg j$ is the result of the (putative) active causal process involving $X = x$ being "shut off" when X is set to the counterfactual $X = x$¢; nor

does it ensure that in $M_{\vec{W}\leftarrow\vec{w}'}$ it is a "causal process" *within* $\vec{Z}$ that produces the change back from $\neg j$ to j when X is returned to its original value $X = x$ in condition C2(b).

The intent of condition C2(a) is to test the dependence of j on $X = x$ by screening off alternative causes for j masking the effect of $X = x$ and the intent of condition C2(b) is to test the ability of $X = x$ to produce j (see Section 3.6). However, because there is no consideration paid to interaction effects in deciding what variables can belong to $\vec{W}$ (i.e., nothing corresponding to $coin_{\vec{u}}(Z; X \mid Y)$) and have their values fixed, it can happen that the setting $X = x'$ in C2(a) or $X = x$ in C2(b), together with the setting $\vec{W} \leftarrow \vec{w}'$, satisfies a minimal sufficient condition (mechanism) for some variable between X and j unsatisfied except for the setting $\vec{W} \leftarrow \vec{w}'$. It is this non-actual mechanism that causes the change from j to $\neg j$ or from $\neg j$ to j, then the Halpern-Pearl definition can lead to counterintuitive causal conclusions. Hopkins and Pearl's (2003) firing squad example (in the context where B does not shoot; see Section 3.8) illustrates the latter case; their first example of an inappropriate G—originally considered by Halpern and Pearl (2000)[18]— illustrates the former:

Example (loanshark) Larry the Loanshark contemplates lurking outside Fred's workplace to cut off his finger, as a warning to him to repay his loan quickly. If Larry cuts off Fred's finger, he will throw it away so that it cannot be reattached. Something comes up, however, so that Larry is not waiting and Larry does not cut off Fred's finger. That same day, Fred has his finger severed by a machine at the factory. He is rushed to the hospital, where the finger is reattached, so if Larry had shown up, he would have missed Fred. At day's end, Fred's finger is functional, which would not have been true had Larry shown up and Fred not had his accident.

Consistent with the lack of consideration paid to structure, neither Halpern and Pearl (2000) nor Hopkins and Pearl (2003) provide structural equations for their models of the scenario. Instead, they provide variables, their values, and leave the structure of the scenario to be inferred from the intuitive truth-functional relations. The following

[18] Halpern and Pearl attribute the origin of the example to Eric Hiddleston.

causal model is consistent with the implied causal model. The propositional endogenous variables, with values 1 (“true”) or 0 (“false”), are

- *LL* for “Larry the loanshark lurks outside Fred’s workplace”;
- *LC* for “Larry cuts off Fred’s finger and throws it away”;
- *A* for “Fred suffers an accident at work where his finger is caught in machinery”;
- *FA* for “Fred’s finger is available”;
- *FS* for “Fred’s finger is severed”;
- *FR* for “Fred’s finger is reattached”; and
- *FF* for “Fred’s finger is functional at day’s end”.

The structural equations are

- $LC = LL \wedge \neg A$;
- $FS = LC \vee A$;
- $FA = \neg LC$;
- $FR = FA \wedge FS$; and
- $FF = FR \vee \neg FS$.

As usual (see Section 3.7), the context $\vec{u}$ is assumed to be such that the endogenous variables have their actual values in the scenario. In this context *LL* = 0 and *A* = 1 so that *LC* = 0 and *FS*, *FA*, *FR*, and *FF* are all equal to 1. The corresponding causal diagram is given in Figure 4.2.

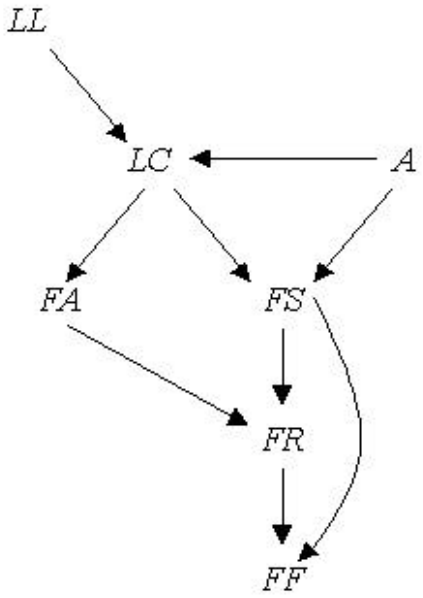

Figure 4.2: Causal Diagram for the Loanshark model.

According to the Halpern-Pearl definition, $FS = 1$ is (counter-intuitively) an actual cause of $FF = 1$; that is, Fred's finger being severed earlier in the day is a cause of his finger being functional later in the day: Letting $\vec{W} = \{LL\}$ and setting $LL = 1$ ($\vec{W} \leftarrow \vec{w}'$) gives $FF = 0$ ($\neg j$) when $A = 0$ ($\vec{X} \leftarrow \vec{x}'$) satisfying condition C2(a). Then returning A to its original value $A = 1$ ($\vec{X} \leftarrow \vec{x}$) in ($M_{LL \leftarrow 1}, \vec{u}$) gives $FF = 1$ (j) and, since every variable in $\vec{Z} = \mathcal{V} - \vec{W}$ returns to its original value in $(M, \vec{u})$, condition C2(b) is satisfied. (Conditions C1 and C3 are trivially satisfied.)

Notice that it is the interaction of the non-actual $LL = 1$ (outside of $\vec{Z}$) with the condition C2(a) counterfactual, $A = 0$, that accounts for $FF = 0$. If the (intuitively) causally irrelevant Larry scenario were absent from the model, or if LL and LC are fixed at their actual values ($LL = 0 \wedge LL = 0$), then FF does not counterfactually depend on FS. Halpern and Pearl (2000) were troubled that the addition of a non-actual, "fanciful" contingency, as in the loanshark example, could change the result of the causal query—in this case, making it counterintuitive. They accepted that if it was a "reasonable possibility" that Larry would show up to cut off and throw away Fred's finger should

Fred appear then it was proper to include the Larry scenario in the model. In that case, even if in fact Larry doesn't show up, according to the structural contingency theory underlying the Halpern-Pearl definition (see Sections 3.6 and 3.7), we are bound to contemplate the interventional contingency that he does; and being so bound, the conclusion that Fred's workplace accident ($FS = 1$) was a cause of his finger functioning at the end of the day ($FF = 1$) is acceptable. Halpern and Pearl (2000) propose to avoid "truly" fanciful scenarios being included in an appropriate model with a calculus of fancifulness—a ranking of contingencies based on the degree of surprise they elicit. As Hopkins and Pearl (2003) show, this is not a very appealing solution[19]:

> Consider what happens if the story is amended such that Larry fully intends to show up at the factory, but is improbably struck by lightning such that he doesn't arrive. Hence the prior probability of LL = 1 is high, and yet we still intuitively would like to conclude that Fred's accident did not cause his finger's functionality. In fact, we would only want to conclude that Fred's accident was a cause of his finger being functional at day's end in the event that Larry shows up in actuality.

Because the possibility of interaction effects, which plagues the Halpern-Pearl definition, does not arise with the new definition, neither does the problem of non-actual contingencies. In the loanshark example, the only routes $\vec{R}$ from A to FF must pass through FA or FS but not both (see Figure 4.2). Since $\neg coin_{\vec{u}}(FR; FS \mid FA)$, if $\vec{R}$ passes through FS then FA will be fixed at its actual value in $M_{[\vec{R},\vec{u}]}$, $FA = 1$ and it is not therefore possible that $FF = 0$ so that condition AC2(a) of the new definition will fail. On the other hand, if $\vec{R}$ passes through FA then $\vec{R}$ must include LC. Since $\neg coin_{\vec{u}}(LC; A \mid LL)$, LL will be fixed at $LL = 0$ in $M_{[\vec{R},\vec{u}]}$ and it is not therefore possible for $FS = 1$ when $A = 0$ so that once again condition AC2(a) will fail. According to the new definition of actual causation, $A = 1$ is not a cause of $FF = 1$. As Hopkins and Pearl require, if Larry does not show up, he is irrelevant to whether the finger functions.

[19] In an unpublished, revised version of their paper, Halpern and Pearl (2002) take a different approach to unreasonable scenarios. They propose to amend the definition of actual causation by permitting the inclusion of an explicit set of disallowed settings to screen out contingencies that tamper with the causal processes to be uncovered.

4.5.2 Definition or Model?

Hopkins and Pearl's (2003) second example of an inappropriate *G* reintroduces Billy and Suzy; this time Billy is up to no good:

Example (bomb) Billy, apparently upset with Suzy's superior rock-throwing ability, places a bomb under Suzy's chair. For some reason Suzy sees fit to look under the chair and notices the bomb. Of course she flees, but is sufficiently undaunted to attend a pre-scheduled medical appointment where she is pronounced healthy.

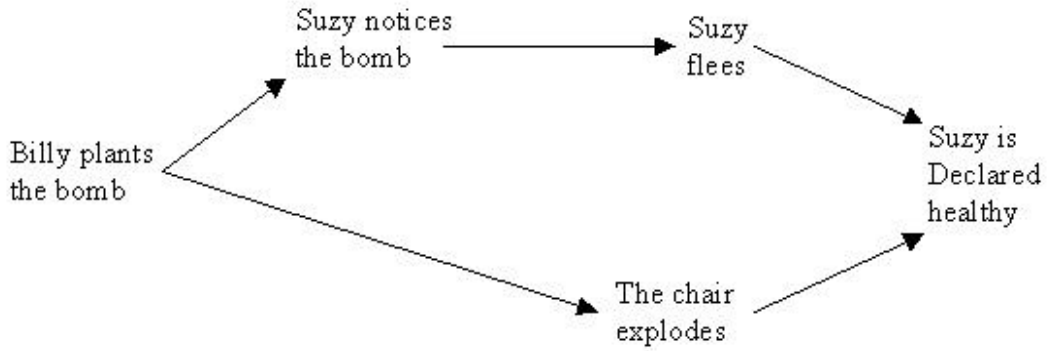

Figure 4.3: Hopkins and Pearl (2003) causal model for the Bomb scenario.

The model provided by Hopkins and Pearl in this case consists of just the diagram in Figure 4.3. Though in this case the structural and (implied) truth-functional equations coincide, it is still useful to flesh out the model in a way that preserves Hopkins and Pearl's point. Let the (propositional) endogenous variables be:

- *BPB* for "Billy plants the bomb;"
- *BBSC* for "There is a bomb beneath Suzy's chair;"
- *SL* for "Suzy looks under the chair;"
- *SNB* for "Suzy notices the bomb;"
- *SF* for "Suzy flees;"
- *CE* for "The chair explodes;"
- *SI* for "Suzy is injured;" and
- *SDH* for "Suzy is declared healthy."

The structural equations are:

- $BBSC = BPB$;
- $SNB = BBSC \wedge SL$;
- $CE = BPB$;
- $SF = SNB$;
- $SI = CE \wedge \neg SF$; and
- $SDH = \neg SI$.

The corresponding causal diagram is given by Figure 4.4.

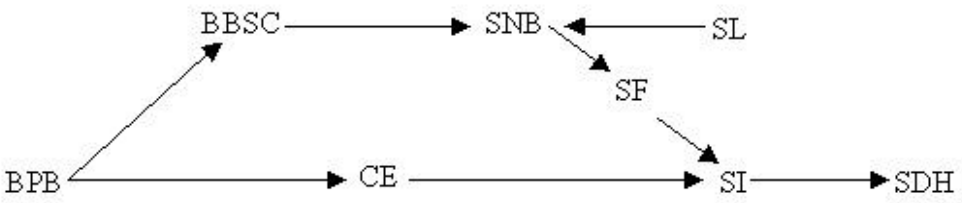

Figure 4.4: Causal diagram for the modified Bomb model.

As Hopkins and Pearl suggest, it seems counterintuitive to classify Billy's planting of the bomb as a cause of Suzy being declared healthy; however, according to the Halpern-Pearl definition it is (for condition C2 let $\vec{W} = \{CE\}$ and set $CE \leftarrow 1$). This is also the case with the new definition since for the path $\vec{R} = \langle BPB, BBSC, SNB, SF, SI, SDH \rangle$ CE must be fixed $CE \leftarrow 1$ in $M_{[\vec{R},\vec{u}]}$ ($\neg coin_{\vec{u}}(SI; SF \mid CE)$) and so condition AC2 will be satisfied.

Hopkins and Pearl assume this result is a problem for the Halpern-Pearl definition and, by implication, for the new definition. However, this assumes that the model described corresponds to the mental model presumed by the structural language (see Section 3.2) to underlie the intuition that Billy's planting of the bomb is not an actual cause of Suzy being declared healthy. In both the original and elaborated models, Billy's bomb planting is not part of the mechanism for Suzy being injured; Billy's misbehaviour is modelled as an indirect participant in any injury suffered by Suzy. In

terms of Pearl's structural contingency theory (see Sections 3.6 and 3.7), the model contemplates the possibility that the chair will explode though no bomb was planted. However, the assumption underlying the intuition that Billy's planting of the bomb does not cause Suzy to avoid injury is that but for the planting of the bomb the chair would not have exploded. Therefore, an equally valid explanation—to the suggestion that there is a problem with the definition—for the counterintuitive result is that the model is inaccurate: in the model underlying the intuition the intermediate variable *CE* between *BPB* and *SI* is factored out by the assumption that the chair does not explode if Billy does not plant the bomb. The "intuitive" model leaves out the variable *CE* and replaces the equation for *SI* with $SI = BPB \wedge \neg SF$ in which case it is not possible to show with either definition that $BPB = 1$ is a cause of $SDH = 1$. (Of course, if the assumption that the chair would not explode if Billy does not plant the bomb is incorrect, then the intuition and revised model are wrong.)

4.6 Actual Causal Subtleties

Hopkins and Pearl (2003) raise some general, "ontological" concerns with respect to any attempt to define actual causation within the structural model framework. While in the previous section it was suggested that Hopkins and Pearl fail to distinguish the separate issues of defective definitions and inaccurate models, here it will be argued that they fail to distinguish between the difficulties in identifying or determining what is an appropriate model for a scenario with the issue of whether the determination of an actual causal query involves the application of a structural definition to a (informal mental or formal) causal model.

Hopkins and Pearl seem to argue that the determination of an actual causal query cannot be merely structural (the application of a counterfactual query to some properly modified structure of events) if there are scenarios about which there are settled causal intuitions but which defy modelling in the structural framework. If there are such scenarios, that would be a problem. Fortunately, a small library of causal structures seems to be able to deal effectively with most scenarios and causal intuitions settled when sufficient information is provided to infer the causal dynamics of the scenario.

Hopkins and Pearl first consider the following example.

Example (magic) The laws of magic require that at midnight the first spell cast the previous day—i.e., since the previous midnight—will take effect. The first spell cast on a given day is by Merlin, who casts a spell to turn the prince into a frog. That evening, Morgana also casts a spell to turn the prince into a frog. At midnight the prince duly becomes a frog.

Hopkins and Pearl (2003) say

> Intuitively, Merlin's spell is a cause of the prince's transformation and Morgana's is not. In this case, although there is preemption, there are no intermediating events that we can really play with and model. Spells work directly, and without Merlin's spell, the prince's transformation would have occurred at precisely the same time and in the same manner. Hence it is far from clear how we could model this story appropriately with a structural model. One concise way to express this story uses first-order constructs. For example, we could neatly express the rule that a spell works iff *there does not exist* [original italics] a previous spell cast that day.

Halpern and Pearl (2002, p. 30) have no difficulty discerning a causal structure for this scenario: "Either spell would have done the job, had it been the only spell of the day; but Merlin's spell was first, so it was his spell that caused the transmogrification. Merlin's spell trumped Morgana's." Halpern and Pearl recognize that this is just the rock-throwing (Section 4.1) or two-switches (Section 4.2) scenario, where there are two causes present in the scenario for the same effect except that one preempts ("trumps") the other: $Frog = Mer \vee (\neg Mer \wedge Mor)$ ($Frog$ = "The prince turns into a frog" and $Mer(Mor) =$ "Merlin (Morgana) casts spell"). The "rule of magic," peculiar as it is, is sufficient to enable the determination of an appropriate causal structure for the scenario, "to ensure that the structural equations properly represent the dynamics of the story" (Halpern and Pearl 2002, p. 30). Formalizing the modelling process, the process by which the information about a scenario is translated into an appropriate causal model, may require a more feature-rich language than the structural language (e.g., to formalize the rule of magic that explains the causal dynamics of the magic scenario), but that is different than arguing that it is more than the causal structural dynamics of the scenario that determines the outcome of an actual causal query. The structural language is adequate to represent the latter.

4.6.1 Temporal Constructs

Hopkins and Pearl (2003) suggest that the structural framework would benefit from temporal constructs, a "fine distinction" difficult to phrase in the structural model framework. They give as an example the difficulty of expressing the difference between an event causing another event and hastening another event as in "a strong wind that causes Suzy's rock to hit the bottle earlier than anticipated, but does not cause the bottle to shatter." It is not clear why such a hastener would be included in a causal model, informal or formal, if it plays no role in the causal dynamics of the scenario. For instance, suppose in the rock-throwing scenario (see Section 4.1) that Billy's rock would have reached and shattered the bottle if the wind had not sped up Suzy's throw so that the latter arrived first. In that case, the wind (W) is a necessary part of the causal dynamics of the scenario:

$$BS = (ST \wedge W) \vee (ST \wedge \neg BT) \vee (\neg ST \wedge BT) \vee (\neg W \wedge BT).$$

However, if Suzy's throw would have reached the bottle first and shattered it whether or not the wind blew her rock there sooner, then there would be a reason for including the wind as part of a causal mechanism for the rock shattering the bottle at a certain time, but there would be no reason for including it in a model for the bottle shattering. Both an informal mental model and a formal model would factor the wind out; this latter modelling process could perhaps benefit from having a means of representing the distinction between genuine causal factors and mere hasteners.

4.6.2 Conditions and Transitions

Hopkins and Pearl (2003) use the example of the difference between a man being dead and a man dying to illustrate the distinction between enduring conditions and transitional events—a distinction they say can be modelled with specialized classes of random variables not available in the structural framework. They suggest the importance of this distinction to the determination of actual causal issues by wondering whether we would be willing to ascribe a heart attack causing the man's death as the cause of him *being dead* in the year 3000. Elsewhere, Hopkins and Pearl (2002) give the example of a man blowing out a candle ($B = 1$) when the wax would have run out in 5 minutes in any case ($WRO = 1$) and ask whether the man blowing out the candle is the cause of the room

being dark one hour later ($RD=1$). The model provided has structural equations $WRO=\neg B$ and $RD=B\vee WRO$. The Halpern-Pearl definition concludes, counter-intuitively, that blowing out the candle is a cause of the room being dark one hour later ($B=1$ is a cause of $RD=1$) even though the candle would have gone out anyway.

First, note that this model does not satisfy coefficient invariance (see Section 4.2): the value of *WRO* depends on the value of *B*. The model confuses the wax running out with there being 5 (or less than 60) minutes of wax left to burn ($WL=1$). A more "appropriate" model would have equations $WRO=\neg B\wedge WL$ and $RD=B\vee WRO$ (see Figure 4.5). However, it is still the case that even for the new definition that $B=1$ is a cause of $RD=1$ (since $coin_{\vec{u}}(RD;B\,|\,WRO)$, the value of *WRO* can be changed to $WRO=0$ in $M_{[\langle B,RD\rangle,\vec{u}]}$, which makes $RD=1$ counterfactually dependent on $B=1$).

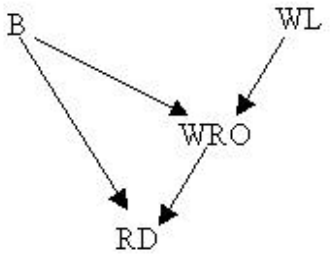

Figure 4.5: Causal diagram for a room being dark?

The model still seems odd because the information provided only describes the causal dynamics for the room *becoming* dark, not for causing the room to stay dark. Intuitively, the cause of an enduring condition is a factor in the absence of which the condition would no longer endure. In the dark room scenario, this would require information that would allow for inferences about lurking mechanisms that would cause the transition from *RD* to $\neg RD$. The same analysis applies to the "being dead" example, where the analogue of *WRO* would be all manners of deaths that could have befallen the heart attack victim but for the heart attack, including old age; though, in the case of being dead, the mechanisms available for becoming not dead are apparently few.

4.6.3 Presence and Absence of an Event

Finally, Hopkins and Pearl (2003) say that in the structural model framework the distinction between the presence and absence of an event is lost. It is not clear that this is the case, at least with the new definition of actual causation. Consider Hopkins and Pearl's firing squad example (see Section 3.8) where the single structural equation is $D = (A \wedge B) \vee C$. Suppose the context $\vec{u}$ is such that $A = B = C = 0$ and therefore $D = 0$. According to the new definition, $A = 0$ (the absence of the event "*A* loads *B*'s gun") is not a cause of $D = 0$ (the absence of the event "the prisoner dies") since $\neg coin_{\vec{u}}(D; A|\, B)$ means effectively $D = C$ in $M_{[\langle A,D\rangle,\vec{u}]}$. ($A = 0$ is a cause of $D = 0$ for the Halpern-Pearl definition.) On the other hand, if the structural equation for *D* was given in the logically equivalent form $\neg D = (\neg C \wedge \neg A) \vee (\neg C \wedge \neg B)$, $A = 0$ will be classified as an actual cause of $D = 0$ (since $coin_{\vec{u}}(D; A|\, B)$, $B \leftarrow 1$ is allowed in $M_{[\langle B,D\rangle,\vec{u}]}$ making $D = 0$ counterfactually dependent on $A = 0$). The significance of the presence or absence of an event is not lost when structure is returned to "structural equations."

5. Formal NESS Sets, Comparisons, and Conclusion

The straightforward interpretation for the NESS test in the structural language, suggested in Section 3.8, fails in the identification of Wright's concept of causal generalizations with the sets $\vec{Z}$ of the Halpern-Pearl definition. The Halpern-Pearl definition fails to formalize the NESS test in a causal world defined by a causal model with a specific context as it can ascribe actual causality to necessary conditions of non-satisfied (non-actual) sets of sufficient conditions. Guided by Hausman and Woodward's (1999) explanation of coefficient invariance, the previous chapter demonstrated how the concept of minimal sufficient sets, identified with component mechanisms in structural equations satisfying what was called "term modularity" in Section 4.2, could be encoded in the structural language. The concept of $coin_{\vec{u}}(Z; X \mid Y)$, along with the dependent concept of $M_{[\vec{R}, \vec{u}]}$, was developed in Sections 4.3 and 4.4 to use that encoded information to isolate an actually active causal process—linked sets of satisfied minimal sufficient conditions (chains of active mechanisms)—along a route. This allowed for a new way of defining an appropriate *G* for the counterfactual strategy (see Section 3.8): depending on its role in a particular equation, a variable can be fixed at its actual value or allowed to vary. Thus, the problem of being too permissive or too strict, a problem for the all-or-nothing choice of appropriate *G* with the Halpern-Pearl approach, is avoided with the new definition of Section 4.4.

This chapter concludes the thesis and is organized as follows. Section 5.1 redefines Halpern and Pearl's "active causal process" (see Section 3.6) using the new definition of actual causation and applies it in comparison with NESS sufficient sets in analysis of cases of duplicative causation (Section 5.2), preemptive causation (Section 5.3), and of double omissions (Section 5.4). Finally, the chapter concludes with an

evaluation of the thesis that the structural language is adequate to formalize the NESS test and suggests directions for future research (Section 5.6).

5.1 An Actually Sufficient Set in a Causal World—NESS Formalized?

Suppose that C is an actual cause of E in $(M,\vec{u})$. Then there exists a route $\vec{R}=\langle C,D_1,\ldots,D_n,E\rangle$ and a sequence of satisfied terms $\vec{T}=\langle T_{D_1},\ldots,T_{D_n},T_E\rangle$ satisfying condition AC2 of the new definition of actual causation. The *active causal process relative to* $\vec{R}$ *and* $\vec{T}$ *in* $\vec{u}$ (denoted $ACP^{\vec{T}}_{[\vec{R},\vec{u}]}$) is the set $\vec{R}\cup\{X_i\}$ where $X_i\in\mathcal{V}-\vec{R}$, $Y\in\vec{R}-E$, $Z\in\vec{R}-A$, and $\neg coin^{\vec{T}Z}_{\vec{u}}(Z;Y\mid X_i)$. That is to say, $ACP^{\vec{T}}_{[\vec{R},\vec{u}]}$ is the subset of variables in $\mathcal{V}$ that have their valued fixed in forming $M^{\vec{T}}_{[\vec{R},\vec{u}]}$ that are parents of a variable in $\vec{R}$. (Again, the term specific terminology and the accompanying superscripts will be discarded where the context does not require them; see Section 4.4.)

By construction of $\vec{T}$, if every variable X in $ACP^{\vec{T}}_{[\vec{R},\vec{u}]}-E$ is set to its actual value (i.e., for all $X\in ACP^{\vec{T}}_{[\vec{R},\vec{u}]}$, $X\leftarrow x$ where $(M,\vec{u})\models(X=x)$), then the resulting set of variable assignments is actual, includes the actual value of C, and is sufficient for the actual value of E. Also, since $\vec{R}$ and $\vec{T}$ satisfy the counterfactual test of condition AC2 of the new definition, the actual value of C is necessary for the sufficiency of the set. Thus, if causal models are interpreted as the causal generalizations that Wright argues define sufficient sets, then the new structural definition of actual causation will classify $C=c$ as an actual cause of $E=e$ only if $C=c$ is a necessary member of a sufficient set of actual conditions. The NESS test is satisfied (formalized) in a causal world.

For example, consider a causal model with the following structural equations:

- $D=A\wedge\neg B$;
- $C=\neg B$; and
- $E=C\vee D$

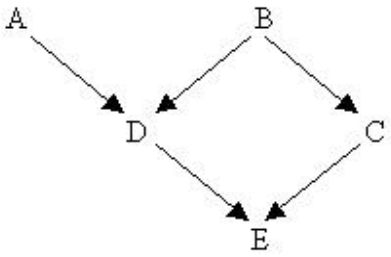

Figure 5.1: Causal diagram for $(M,\vec{u})$

The causal diagram for this model is represented in Figure 5.1. Assume that $\vec{u}$ is such that $A=1$ and $B=0$, and therefore $C=D=E=1$. To show that $A=1$ is an actual cause of $E=1$, note that condition AC1 is satisfied and let $\vec{R}=\langle A,D,E\rangle$ for condition AC2. Since $coin_{\vec{u}}(E;D|\,C)$ and $\neg coin_{\vec{u}}(D;A|\,B)$, the structural equations in $M_{[\vec{R},\vec{u}]}$ are $D=A$, $C=\neg B$, and $E=C\vee D$. The corresponding causal diagram is represented in Figure 5.2. Letting $\vec{W}=(B,C)$ and setting $\vec{W}=\vec{w}=(0,0)$ satisfies conditions AC2(a) and (b).

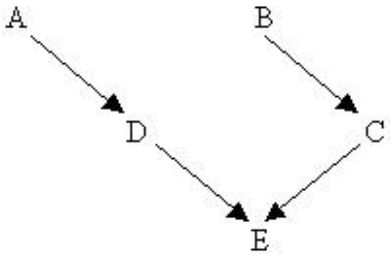

Figure 5.2: Causal diagram for $M_{[\vec{R},\vec{u}]}$

$ACP_{[\vec{R},\vec{u}]}$ tells a complete causal story for $E=1$ in the scenario modelled by $(M,\vec{u})$; it says if $A=a$, $B=b$, and $D=d$, where $(M,\vec{u})\models(A=a\wedge B=b\wedge D=d)$, then $E=e$ irrespective of what happens in the rest of the causal world $(M,\vec{u})$. The remaining links into E in Figure 5.2 (the route $\langle B,C,E\rangle$) could be removed without effect. $\{A=a,B=b,D=d\}$ is a sufficient set of conditions to guarantee $E=e$.

Condition AC2 says that $A = a$ is a necessary condition for that set. In other words, the actual values of $ACP_{[\vec{R},\vec{u}]} - E$ define a NESS set for $E = e$, an actually sufficient set, in the causal world $(M, \vec{u})$. In the above example, $A = 1$ is a necessary condition for the sufficiency of the set $\{A = 1, B = 0, D = 1\}$. It is also possible to give meaning to Wright's "omnibus negative condition" (see Section 2.3.5): the variables in $ACP_{[\vec{R},\vec{u}]} - \vec{R}$ are the potentially preemptive variables whose values must be checked to ensure the "absence of preventing or counteracting causes." Thus the variables not in $ACP_{[\vec{R},\vec{u}]}$ can be ignored in telling the causal story (the NESS set)—as Wright does in describing the NESS sets in duplicative causation cases (see Section 5.2)—the variables in $ACP_{[\vec{R},\vec{u}]}$ must be included in telling the causal story in preemption cases.

5.2 Preemptive Causation Cases

Wright (1985, p 1795) considers two preemptive scenarios: in the first, *D* shoots and kills *P* before *P* can drink tea fatally poisoned by *C* and, in the second, *D* shoots and instantly kills *P* after *P* drinks tea fatally poisoned by *C* but before the poison can take effect. With respect to the first scenario (poisoned tea), in Wright's (1985, p 1795) NESS analysis,

> D's shot was necessary for the sufficiency of a set of actual antecedent conditions that did not include the poisoned tea. Conversely, C's poisoning of the tea was not a necessary element of any sufficient set of actual antecedent conditions. A set that included the poisoned tea but not the shooting would be sufficient only if P actually drank the tea, but this was not an actual condition. The shooting preempted the potential causal effect of the poisoned tea.

In this scenario, the story of death by poisoning would have occurred (the intake of the poison through consumption of the tea will have occurred) but for *D* shooting *P*. This is reflected in the following causal model. The model has the following propositional variables:

- *DS* represents "*D* shoots;"
- *PT* represents "*C* poisons the tea;"

- CP represents "P consumes poison;" and
- PD for "P dies."

The structural equations are:

- $CP = \neg DS \wedge PT$ and
- $PD = DS \vee CP$.

The causal diagram corresponding to these equations is represented in Figure 5.3.

To show that $DS = 1$ is an actual cause of $PD = 1$, let $\vec{R} = \langle DS, PD \rangle$ for condition AC2. Since $coin_{\vec{u}}(PD; DS \mid CP)$, $M_{[\vec{R},\vec{u}]} = (M, \vec{u})$ and therefore $\vec{W} = (PT, CP)$. Setting $\vec{W} = (0,0)$ then satisfies conditions AC2(a) and (b). Note that $ACP_{[\vec{R},\vec{u}]} = \vec{R}$ and the NESS set including $DS = 1$ for $PD = 1$ in $(M, \vec{u})$ is just $\{DS = 1\}$, as it is with Wright's analysis.

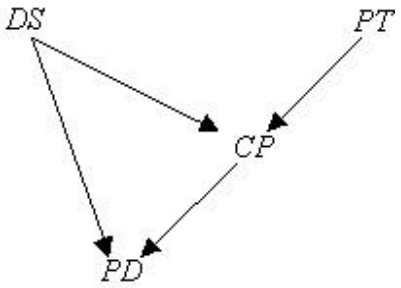

Figure 5.3: Causal diagram for the poisoned tea scenario

Suppose that the context was such that D does not shoot ($\neg DS$) but P still poisons the tea. Then $CP = 1$ and $PD = 1$ and it is straightforward to show that $PT = 1$ is a cause of $PD = 1$ by letting $\vec{R} = \langle PT, CP, PD \rangle$ in condition AC2. Note, however, that since $\neg coin_{\vec{u}}(CP; PT \mid DS)$, $ACP_{[\vec{R},\vec{u}]} = \{PT, CP, DS, PD\}$ and the NESS set including $PT = 1$ for $PD = 1$ in $(M, \vec{u})$ is $\{PT = 1, CP = 1, DS = 0\}$: the absence of the preempting condition DS must be included.

For the second example, Wright's (1985, p 1795) NESS analysis of why *D*'s shooting was a cause of *P*'s death is the same as that for the first example; as to whether *C*'s poisoning of the tea was a cause:

> Even if P actually had drunk the poisoned tea, C's poisoning of the tea still would not be a cause of P's death if the poison did not work instantaneously but the shot did. The poisoned tea would be a cause of P's death only if P drank the tea and was alive when the poison took effect. That is, a set of actual antecedent conditions sufficient to cause P's death must include poisoning of the tea, P's drinking the poisoned tea, and P's being alive when the poison takes effect. Although the first two conditions actually existed, the third did not. D's shooting P prevented it from occurring. Thus, there is no sufficient set of actual antecedent conditions that includes C's poisoning of the tea as a necessary element. Consequently, C's poisoning of the tea fails the NESS test. It did not contribute to P's death.

A causal model for this scenario differs from the previous one by the addition of a variable PTE for "the poison takes effect." The structural equation for PD becomes $PD = DS \vee PTE$ and the equation for CP becomes $CP = PT$ (see Figure 5.4). As with Wright's NESS analysis, the proof that $DS = 1$ is an actual cause of $PD = 1$ would be essentially the same as with the previous example. The interesting aspect of this example is that it shows how the NESS test allows Wright to "solve" the causal issue by fitting it into the preemptive scenario, illustrating again that the difficult aspect of actual causal queries is deriving the correct model. Given the model, the determination of the actual causal query is straightforward.

5.3 Duplicative Causation Cases

Among the duplicative causation cases, of particular interest are a group of pollution cases where defendants were found liable though none of their individual acts (their "contributions" to the pollution) was sufficient, or necessary given the contributions of the other defendants, to produce the plaintiff's injuries (some adverse effect on the use of his property).[20] Wright (1985, p 1793) applies the NESS test to an

[20] For example, Wright (2001, p 1100) cites the case of Warren v. Parkhurst, 92 N.Y.S. 725 (N.Y. Sup. Ct. 1904), *aff'd*, 93 N.Y.S. 1009 (A.D.1905), *aff'd*, 78 N.E. 579 (N.Y. 1906), where each of twenty-six defendants discharged "nominal" amounts of sewage into a creek which individually were not sufficient to destroy the use of downstream plaintiff's property but the stench of the combined discharges was sufficient.

idealized example in which, five units of pollution are necessary and sufficient for the plaintiff's injury and seven defendants discharge one unit each. The NESS test requires only that a defendant's discharge be necessary for the sufficiency of *a* set of actual antecedent conditions, and that (Wright 1985, p 1795)

> Each defendant's one unit was necessary for the sufficiency of a set of actual antecedent conditions that included only four of the other units, and the sufficiency of this particular set of actual antecedent conditions was not affected by the existence of two additional duplicative units.

In fact, in this sense, for each defendant's discharge there are fifteen distinct actually sufficient sets of antecedent conditions, one for each possible choice of any four of the 6 remaining defendant's units of pollution.

The causal model for this example has variables X_i, $i=1,\ldots,7$, representing whether defendant *i* contributed his one unit of pollution $(X_i=1)$ or not $(X_i=0)$. The single structural equation

$$DP=(X_1=1\wedge X_2=1\wedge X_3=1\wedge X_4=1\wedge X_5=1)\vee\ldots\vee$$
$$(X_7=1\wedge X_6=1\wedge X_5=1\wedge X_4=1\wedge X_3=1)$$

consists of 21 terms where each term is a conjunction of 5 of the 7 literals $X_i=1$. Since each literal $X_i=1$ is satisfied in the given scenario $(M,\vec{u})$, each literal occurs in 15 satisfied terms in conjunction with 4 of the remaining 6 X_i or, equivalently, each literal $X_i=1$ occurs in 15 terms without conjuncts involving 2 of the remaining 6 variables. Thus, for any $X_i=1$ and variables X_k, X_l $(i\neq k\neq l)$, there exists some term T_{DP} with $coin_{\vec{u}}^{T_{DP}}(DP; X_i \mid X_k, X_l)$.

Without loss of generality, to show that each defendant's pollution discharge is an actual cause of $DP=1$, let $i=1$ and choose T_{DP} so that $coin_{\vec{u}}^{T_{DP}}(DP; X_1 \mid X_6, X_7)$ (i.e., $T_{DP}=(X_1=1\wedge X_2=1\wedge X_3=1\wedge X_4=1\wedge X_5=1)$). Then with $\vec{R}=\langle X_1, DP\rangle$ and $\vec{T}=\langle DP\rangle$, the equation for *DP* in $M_{[\vec{R},\vec{u}]}^{\vec{T}}$ is

$$DP = (X_1 = 1) \vee (X_6 = 1) \vee (X_7 = 1) \vee (X_1 = 1 \wedge X_6 = 1) \vee (X_1 = 1 \wedge X_7 = 1) \\ \vee (X_6 = 1 \wedge X_7 = 1) \vee (X_1 = 1 \wedge X_6 = 1 \wedge X_7 = 1)$$

Since, in $M^{\vec{T}}_{[\vec{R},\vec{u}]}$, *DP* is a trivial function of the variables in $\{X_2,\ldots,X_5\}$, for $\vec{W} = \mathcal{V} - \vec{R} = \{X_i\}$, $i = 2,\ldots,7$, of condition AC2 of the new definition, only the settings for X_6 and X_7 matter. Setting $(X_6, X_7) = (0,0)$, $X_1 = 1$ is easily seen to satisfy the counterfactual test of conditions AC2.

Note that $ACP^{\vec{T}DP}_{[\vec{R},\vec{u}]} = \{X_1, X_2,\ldots, X_5, DP\}$ and, in the causal world $(M,\vec{u})$, $X_1 = 1$ is necessary for the sufficiency of the set including defendant one's discharge ($X_1 = 1$) and only four other discharges.

5.4 Double Omission Cases

Recall (see Section 2.3.5) a class of cases that proved difficult for Wright's NESS test analysis, the so-called "double omissions cases." At this point in the thesis the suspicion should arise that the difficulty in these cases is likely one of modelling, and that is the case.

Wright's (1985, p. 1801) analysis of the braking case described in Section 2.3.5 is reproduced here for convenience:

> It is clear that *D*'s negligence was a preemptive cause of *P*'s injury, and that *C*'s negligence did not contribute to the injury. *D*'s failure to try to use the brakes was necessary for the sufficiency of a set of actual antecedent conditions that did not include *C*'s failure to repair the brakes, and the sufficiency of this set was not affected by *C*'s failure to repair the brakes. A failure to try to use brakes will have a negative causal effect whether or not the brakes are defective. On the other hand, *C*'s failure to repair the brakes was not a necessary element of any set of antecedent actual conditions that was sufficient for the occurrence of the injury. Defective brakes will have an actual causal effect only if someone tries to use them, but that was not an actual condition here. The potential negative causal effect of *C*'s failure to repair the brakes was preempted by *D*'s failure to try to use them.

The causal model implied by this analysis has variables:

- *RB* for "repairs brakes;"

- AB for “applies brakes;”
- BO for “brakes operate;” and
- PH for “pedestrian is hit.”

The question is, what are the structural equations? Recall it was argued against the NESS analysis that the roles of C and D are symmetrical and that a NESS argument for one’s negligent act being an actual cause of the pedestrian being hit would work equally well as argument for the other’s negligent act. The structural equations suggested by this might be

- $\neg BO = \neg RB \vee \neg AB$ and
- $PH = \neg BO$.

It is easy to see that the new definition will classify both $\neg RB$ and $\neg AB$ as actual causes of PH for this model. On the other hand, suppose the structural equations are

- $BO = RB \wedge AB$ and
- $PH = \neg BO$.

In that case the new definition will classify neither RB nor AB as a cause of PH. This model captures the intuition that not repairing the brakes is not a cause of the pedestrian being hit if the brakes are not applied but also suggests that not applying the brakes cannot cause the striking of the pedestrian if the brakes are not operative. But notice that in Wright’s analysis there is the suggestion of a mechanism that neither of these models includes: “A failure to try to use brakes will have a negative causal effect whether or not the brakes are defective.” In other words, there are two distinct mechanisms for the pedestrian being hit; the confusion arises because not braking just happens to play a part in both. The causal model matching Wright’s original model would have equations

- $BO = RB \wedge AB$ and
- $PH = \neg BO \vee \neg AB$.

The causal diagram for this model is represented in Figure 5.4.

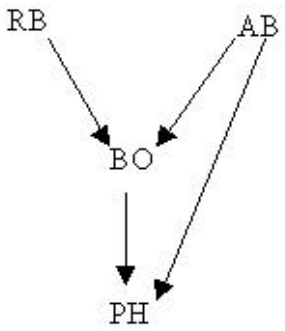

Figure 5.4: Causal diagram for the braking scenario

Indeed, for this model the new definition will classify $\neg AB$ as a cause of PH but not $\neg RB$. It is this missing mechanism that lies behind the intuitive and analytic confusion in the double omission cases. The double omission cases illustrate again that the difficult actual causal issues arise when the causal dynamics of the scenario are not clear or not understood. Once the causal dynamics are understood, the determination of the actual causal query is mechanical.

5.5 Conclusion and Future Work

If the counterfactual strategy for defining actual causation in Pearl's structural language is to succeed, the rules controlling which variables may have their values altered must be cognizant of the mechanisms (the sufficient sets) to which the variables belong and be able to treat variables differently depending on the roles they play as direct causes of dependent variables. Variables may have to be fixed at their actual values in the equation for some variables to avoid activating non-actual causal processes and also have there values set at non-actual values in the equation for some other variables to screen an actual causal process from other causal processes that mask the counterfactual dependency in cases of overdetermination. Previous approaches failed or were limited in their application because they did not observe these requirements.

A definition, such as the one developed in this thesis, which satisfies the first requirement, recognizes the intuition that underlies the NESS test. Not only is it possible to adequately represent the NESS test in Pearl's structural language, as has been done here, it is likely required for a successful definition of actual causation: the limitations of

the Halpern-Pearl definition result from its insensitivity to the relation among variables of being elements of sufficient conditions.

A formal representation of the NESS test in the structural language avoids the conceptual contradictions inherent in Wright's exposition of the concept of the NESS test. However, as Halpern and Pearl emphasized (see Section 3.7), the outcome of an actual causal query still will depend on the accuracy, or appropriateness, of the causal model employed to represent the scenario under consideration. Indeed, the application of the definition developed in this thesis is limited to scenarios whose causal structure can be faithfully modelled by causal models exhibiting what is called here "term modularity."

Thus, among the issues to be explored in the future is whether there are scenarios, as Hopkins and Pearl have suggested (see Section 4.6), that cannot be adequately modeled in the structural language, in particular, cannot be represented by models satisfying term modularity; or whether there are general principles concerning the nature of actual causal queries that rule out such scenarios. (It could be that the ability of individuals to comprehend a scenario in terms of actual causal relations requires the scenario to be internally represented in a way that is naturally amenable to term-modular causal modelling.)

Finally, if the approach developed in this thesis is to be of practical use in artificial intelligence (for example, as a means of providing autonomous intelligent agents with a concept of singular causation) then it needs to be treated algorithmically and its computational complexity characteristics determined. Even then, to be useful, there have to be models. Where those models come from and how the choice of model relates to the context of the inquiry (recall the bomb example of Section 4.5.2 where two apparently valid models of the same scenario resulted in different answers to the actual causal query) are issues that ultimately have to be understood for the definition to be useful. These latter issues "move the issue to the right arena" (Halpern and Pearl 2000, p. 2; see Section 3.7), causal modelling.

REFERENCES

Dawid, A.P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95, pp. 407-448.

Druzdzel, M.J., and Simon, H.A. (1993). Causality in Bayesian belief networks. In Heckerman, D. and Mamdani, A. (Eds.), *Proceeding of the 9th Conference on Uncertainty in Artificial Intelligence*, pp. 3-11. San Mateo, CA: Morgan Kaufmann.

Eiter, T., and Lukasiewicz, T. (2001). Complexity results for structure-based causality. In Proceedings of the International Joint Conference on Artificial Intelligence, pp. 35–40. San Francisco, CA: Morgan Kaufmann.

Fumerton, R.A., and Kress, K. J. (2001). Causation and the law: Preemption, lawful sufficiency, and causal sufficiency [Electronic version]. *Law and Contemporary Problems*, 64 (4), pp. 83-105.

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

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

Hall, N. (2003) Two concepts of causation. Manuscript. To appear in Collins, J., Hall, N., and Paul, L. (Eds.), *Causation and Counterfactuals*. (M.I.T. Press, 2003).

Halpern, J.Y., and Pearl, J. (2000). Causes and explanations: a structural-model approach. Retrieved September 3, 2001 from http://www.cs.cornell.edu/home/halpern/topics.html#rau (Part I, Causes, appears in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, 194-202, 2001.)

Halpern, J.Y., and Pearl, J. (2002). Causes and explanations: a structural-model approach. Part I: Causes. Retrieved September 12, 2002 from the arXiv.org e-Print archive: http://arxiv.org/abs/cs.AI/0011012

Hart, H.L.A., and Honoré, A.M. (1985). *Causation in the law* (2nd ed.). Oxford University Press.

Hausman, D.M., and Woodward, J. (1999). Independence, Invariance and the Causal Markov Condition. British Journal for the Philosophy of Science, 50, pp. 521-583.

Honore, A.M. (2001). Causation in the law. *The Stanford Encyclopedia of Philosophy* (Summer 2002 Edition), Edward N. Zalta (Ed.), URL=http://plato.stanford.edu/archives/sum2002/entries/causation-law/

Hopkins, M. A Proof of the Conjunctive Cause Conjecture in "Causes and Explanations: A Structural Model Approach." UCLA Cognitive Systems Laboratory, Technical Report R-304, 2002.

Hopkins, M., and Pearl, J. (2002). Causality and Counterfactuals in the Situation Calculus. UCLA Cognitive Systems Laboratory, Technical Report R-301, 2002.

Hopkins, M., and Pearl, J. (2003) Clarifying the Usage of Structural Models for Commonsense Causal Reasoning. In Proceedings of the 2003 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University, March 24-26, 2003.

Kim, J. (1993). Causes and events: Mackie on causation. In (Sosa and Tooley, 1993). (Reprinted from *Journal of Philosophy*, 68 (1971), pp. 426-441.)

Lewis, D. (1979). Counterfactuals Dependence and Time's Arrow. Nous, 13, pp. 455-76.

Lewis, D. (1993). Causation. In (Sosa and Tooley, 1993). (Reprinted from *Journal of Philosophy*, 70 (1973), pp. 556-567.)
Mackie, J.L. (1974). *The Cement of the Universe*. Oxford University Press.
Mackie, J.L. (1993). Causes and conditions. In (Sosa and Tooley, 1993). (Reprinted in part from *American Philosophical Quarterly*, 2 (4) (1965), pp. 245-264.)
McCarthy, J. (1990). *Formalizing common sense: Papers by John McCarthy*. (Lifschitz, V., ed.). Norwood, N.J.: Ablex Pub. Corp.
Menzies, P. (2001). Counterfactual theories of causation. *The Stanford Encyclopedia of Philosophy* (Summer 2002 Edition), Edward N. Zalta (Ed.), URL=http://plato.stanford.edu/archives/sum2002/entries/causation-counterfactual/
Ortiz, C.L., Jr. (1999). Explanatory update theory: Applications of counterfactual reasoning to causation. *Artificial Intelligence*, 108, pp. 125-178.
Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82 (4), 669–710.
Pearl, J. (1998). On the definition of actual cause. Technical Report (no. R-259), Department of Computer Science, University of California, Los Angeles.
Pearl, J. (1999). Reasoning with cause and effect. Technical Report (no. R-265), Department of Computer Science, University of California, Los Angeles. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA 1437-1449, 1999.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.
Robertson, D.W. (1997). The common sense of cause in fact. *Texas Law Review*, 75, pp. 1765-1800.
Simon, H.A. (1953). Causal ordering and identifiability. In Hood, W. C., and Koopmans, T. C. (Eds.), *Studies in econometric method*, pp. 49-74. New York: Wiley.
Simon, H.A. (1969). *The sciences of the artificial.* Cambridge : M.I.T. Press.
Sosa, E., and Tooley, M. (Eds.) (1993). *Causation* (Oxford Readings in Philosophy). Oxford University Press.
Woodward J. (2000). Explanation and Invariance in the Special Sciences. British Journal for the Philosophy of Science, 51, pp. 197-254.
Woodward, J. (2002). What is a Mechanism? A Counterfactual Account. Philosophy of Science, 69, pp. S366-S377.
Wright, R.W. (1985). Causation in tort law. *California Law Review*, 73, pp. 1735-1828.
Wright, R.W. (1988) Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa Law Review*, 73, pp. 1001-1077.
Wright, R.W. (2001). Once more into the bramble bush: Duty, causal contribution, and the extent of legal responsibility [Electronic version]. *Vanderbilt Law Review*, 54 (3), pp. 1071-1132.
Yablo, S. (2003) Advertisement for a Sketch of an Outline of a Proto-Theory of Causation. To appear in Collins, J., Hall, N., and Paul, L. (Eds.), Causation and Counterfactuals. (M.I.T. Press, 2003).