

Terms of Use

The copyright of this thesis is owned by its author. Any reproduction, adaptation, distribution or dissemination of this thesis without express authorization is strictly prohibited.

All rights reserved.

CAUSAL MODELING, REVERSIBILITY,
AND LOGICS OF COUNTERFACTUALS

LAM WAI YIN

MPHIL

LINGNAN UNIVERSITY

2012

CAUSAL MODELING, REVERSIBILITY,
AND LOGICS OF COUNTERFACTUALS

by
LAM Wai Yin

A thesis
submitted in partial fulfillment
of the requirements for the Degree of
Master of Philosophy in Philosophy

Lingnan University

2012

ABSTRACT

Causal Modeling, Reversibility, and Logics of Counterfactuals

by

LAM Wai Yin

Master of Philosophy

This thesis studies Judea Pearl's logic of counterfactuals derived from the causal modeling framework, in comparison to the influential Stalnaker-Lewis counterfactual logics.

My study focuses on a characteristic principle in Pearl's logic, named reversibility. The principle, as Pearl pointed out, goes beyond Lewis's logic. Indeed, it also goes beyond the stronger logic of Stalnaker, which is more analogous to Pearl's logic. The first result of this thesis is an extension of Stalnaker's logic incorporating reversibility. It will be observed that the translation of reversibility from Pearl's language to the standard language for conditional logic deserves some attention. In particular, a straightforward translation following Pearl's suggestion would render reversibility incompatible with Stalnaker's logic. A new translation of reversibility will be proposed, and an extension of Stalnaker's logic with the inclusion of the translated reversibility will be investigated. More importantly, it will be shown that the extended Stalnaker's logic is sound and complete with respect to a modified Stalnaker's semantics.

The extension of Stalnaker's logic has an interesting implication. Zhang, Lam, and de Clercq (2012) have shown that a special case of reversibility, despite its name, actually states an important kind of irreversibility: counterfactual dependence (as defined by David Lewis) between distinct events is irreversible. In other words, reversibility entails that there is no cycle of counterfactual dependence featuring two distinct events. This thesis extends their result with respect to different generalizations of reversibility. However, as shown in Zhang et al. (2012), Pearl's logic does not rule out cycles of counterfactual dependence altogether. It in fact allows cycles that involve three or more distinct events. This is peculiar because the status of cyclic counterfactual dependence seems no more metaphysically secure than that of mutual counterfactual dependence. This consideration leads to an exploration of logics that rule out all cycles of counterfactual dependence. A surprising result is that the extension of Stalnaker's logic is precisely a logic of this sort.

DECLARATION

I declare that this is an original work based primarily on my own research, and I warrant that all citations of previous research, published or unpublished, have been duly acknowledged.

(Lam Wai Yin)

6th September, 2012

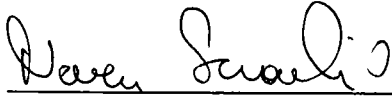
CERTIFICATE OF APPROVAL OF THESIS

CAUSAL MODELING, REVERSIBILITY,
AND LOGICS OF COUNTERFACTUALS

by
LAM Wai Yin

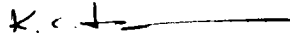
Master of Philosophy

Panel of Examiners :



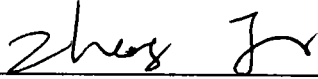
(Chairman)

Prof. SESARDIC Neven



(External Member)

Prof. CHEUNG Kam Ching, Leo



(Internal Member)

Dr. ZHANG Jiji



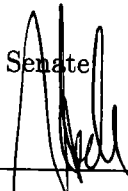
(Internal Member)

Dr. DE CLERCQ Rafael

Chief Supervisor :

Dr. ZHANG Jiji

Approved for the Senate



Prof. SEADE Jesús

Chairman, Postgraduate Studies Committee

Date

CONTENTS

| | |
|---|----|
| Acknowledgements..... | ii |
| Chapter 1. Introduction..... | 1 |
| Chapter 2. Incorporating the Principle of Reversibility in Stalnaker’s Logic..... | 4 |
| 2.1. Similarity-based Theories of Counterfactuals..... | 4 |
| 2.2. Pearl’s Structured-based Theory of Counterfactuals... | 8 |
| 2.3. Reversibility and the Problem of Translation..... | 15 |
| 2.4. Incorporating Reversibility in Stalnaker’s Logic – VCSR | 19 |
| Chapter 3. Pearl’s Reversibility and the Irreversibility of Counterfactual Dependence..... | 27 |
| 3.1. Counterfactual Dependence and its Irreversibility..... | 27 |
| 3.2. Pearl’s Reversibility and the Irreversibility of Counterfactual Dependence..... | 29 |
| 3.3. Cyclic Counterfactual Dependence and a Peculiarity in Pearl’s Logic..... | 41 |
| Chapter 4. Conclusion..... | 50 |
| Appendix..... | 52 |
| Bibliography..... | 54 |

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support of many people. Firstly, the special thank goes to my very helpful advisor, Dr. Jiji Zhang. His enthusiastic supervision helps me explore the realm of causality in philosophy, and his guidance leads me to achieve rigor and professionalism in doing analytic philosophy.

I am also highly indebted to many of my teachers, including Prof. Leo Cheung, Dr. Rafael de Clercq, Prof. Paisley Livingston, Dr. Alex Lo, Prof. Neven Sesardic, Dr. Kelly Trogon, Dr. Yujian Zheng, and Dr. Lei Zhong, among many others. Their support and advice are indispensable for the completion of this thesis. Their comments, especially those from the members of the Panel of Examiners, are undoubtedly precious. In addition, I wish to express my gratitude towards Prof. James Woodward for his participation in my research seminar and his invaluable suggestion on my research.

I greatly appreciate the assistance of my department as well. Sincere help from the staff members contributes a lot to my study in all these years. Moreover, my grateful thanks also go to Mr. Wayne Smith from the CEAL Writing Tutorial Service. His efficient editing gives me a big hand.

Last but not least, I would like to thank my parents, my sister and Ms. Yuuka Yeung. My effort mainly results from their persistent support. There is no doubt that they are the main contributors to my work. I hereby dedicate this thesis to them with my gratitude.

CHAPTER 1

Introduction

Consider the following conditional statement: *had* Barack Obama lost in the election to the U.S. presidency in 2008, he *would not have been* the 2009 Nobel Peace Prize laureate. The conditional has a contrary-to-fact antecedent because, in fact, Obama did win the election in 2008. It is a received view in philosophy that there are two different kinds of conditionals, *indicative* and *subjunctive*.¹ My primary concern in this thesis is the subjunctive kind, or *counterfactual conditionals*, which are often expressed in a subjunctive mood with auxiliary verbs such as “were”, “had” and “would”, like the example just stated.

Counterfactual reasoning is common. Indeed, we always wonder what things would turn out to be were things to happen differently, or what would happen in a counterfactual possibility. Academically, counterfactual reasoning has drawn the attention from different disciplines.² In particular, philosophers and logicians study the question of whether counterfactual conditional statements are truth bearers, and, if they are, inquire the general truth condition of them.

The scrutiny of the truth condition and the assertability condition of conditional statements in general can be traced back to its contemporary philosophical root in Ramsey (1990/1929). The discussion of the issue later was developed more acutely by Chisholm (1946), and especially by Goodman (1955) among other innumerable writers.³ The study of *cotenability* advocated by Goodman, generally speaking, dominates the discussion of counterfactual conditionals approximately from 1950s to 1960s. To give a basic idea, a counterfactual conditional “if P were the case, Q would be the case” is true, according to the cotenability theorists, if and only if Q deductively follows from P together with some set of facts and laws of nature.

The cotenability theory ceased to be the paradigm after certain serious criticisms prevailed, especially those aimed at the unclear constraints on the sets of

¹See Bennett (2003) for an in-depth survey on the indicative/subjunctive distinction.

²For example, Weber (1949) has argued for the significance of counterfactual inquiry in historical explanation in sociology. Elster (1978) and Fearon (1991) have embraced the counterfactual strategy in their research of political science. Psychologists, for instance, Kahneman and Miller (1986), have also admitted the importance of counterfactual alternatives in the study of norms. See Hendrickson (2010) for a detailed literature review on how counterfactual reasoning has been broadened to diversified disciplines.

³See the entry of *The Logic of Conditionals* in the Stanford Encyclopedia of Philosophy (SEP) for a history of the discussion.

facts. Subsequently, a more prominent project came to take over the counterfactual analysis, which is referred as the *world-similarity-based*, or *similarity-based* for short, theories of counterfactuals. The majority of similarity-based theories are attributed to the work of Robert Stalnaker (1968) and David Lewis (1973b). They both make use of the possible-world semantics to study counterfactual possibilities. The truth of a counterfactual conditional depends on whether the consequent holds in the possible world(s) most similar to the actual world which satisfies the counterfactual antecedent. However, to determine whether a world w is more similar, or *closer*, than another world w' to our world is not easily demonstrated. Lewis (1979) developed a system of measuring instruments which enable philosophers to determine the distance between possible worlds. The discussion on his system of measurement later became a focal point for philosophers to assess the similarity-based theories of counterfactuals.⁴

Lewis (1973a) also famously proposed a counterfactual theory of causation, based on his theory of counterfactuals. The classical definition of causation given by Hume (1751/1748) is popularly referred to the following two statements of his - "...we may define a cause to be 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 had not been, the second never had existed". Let alone the relation in Hume's definition of causation, Lewis shed light on the second statement formulated in a counterfactual conditional whereas his contemporaries (e.g. Armstrong (1978)) focused on criticizing the first statement concerning regularity. Even though Lewis's counterfactual theory of causation did receive criticisms, Lewis's contribution to the discussion of causation is beyond any doubt.⁵

More recently, researchers have facilitated the study of causality with rigorous mathematical tools borrowed from econometricians (e.g. Strotz and Wold 1969; Fisher 1970). Instead of the reductive analysis, they favor the non-reductive account of causation by the means of *intervention*, which is essentially a causal term (e.g. Halpern (2000), Hausman (2005), Hitchcock (2001, 2007), Pearl (1995, 2009), Spirtes, Glymour, and Scheines (2000), Woodward (2003)).

The interventionists, particularly Halpern and Pearl, use causal models to give accounts of both (actual) causation and counterfactuals. That is, the counterfactual conditional "had the variable X been the value x , the variable Y would have taken the value y " is true if and only if by appropriately intervening X to be x , Y would take the value y .

⁴For some examples, see Bennett (1984), Bigaj (2004), Elga (2001), Field (2003), Hausman (1998), Hiddleston (2005).

⁵See Paul (2009) for an overview of the counterfactual theory of causation.

The interventionists do realize the influence of Lewis’s theory of causation in philosophy and the similarity between their work and Lewis’s. Hence, almost every of the mentioned interventionists has compared their own work to Lewis’s. In this thesis, I focus on a characteristic principle in Pearl’s logic, named reversibility. The general layout of the remaining chapters is as follows.

Chapter 2, firstly, offers an overview of the similarity-based theories of counterfactuals from both Lewis and Stalnaker, and also the structure-based theory of counterfactuals originated by Pearl. It will be argued that the principle of reversibility does not only go beyond Lewis’s logic, but also goes beyond Stalnaker’s logic, which is more analogous to Pearl’s than Lewis’s in light of Pearl’s principle of definiteness. The first ambition of this thesis is to extend Stalnaker’s logic to incorporate reversibility. It will be observed that the translation of reversibility from Pearl’s language to the standard language for conditional logic deserves certain attention. Particularly, a straightforward translation following Pearl’s suggestion would render reversibility incompatible with Stalnaker’s logic. An appropriate translation will be proposed, and an extension of Stalnaker’s logic with the inclusion of the translated reversibility will be investigated. Next, Stalnaker’s semantics will be modified and the soundness and completeness of the extension will be demonstrated with respect to the modified Stalnaker’s semantics.

Another important result of this thesis is presented in chapter 3, which extends the work from Zhang, Lam, and de Clercq (2012). We revealed that a special case of reversibility, despite its name, actually states an important kind of irreversibility: counterfactual dependence (as defined in Lewis (1973a)) between distinct events is irreversible. Putting it differently, reversibility entails that there is no cycle of counterfactual dependence featuring two distinct events. This result will be extended with respect to different generalizations of reversibility. Curiously, however, we have also shown that Pearl’s logic does not rule out cycles of counterfactual dependence altogether. It in fact allows cycles that involve three or more distinct events. From a metaphysical perspective, the status of cyclic counterfactual dependence seems no more secure than that of mutual counterfactual dependence. This consideration leads me to explore logics that rule out all cycles of counterfactual dependence. A main result is that the extension of Stalnaker’s logic I developed in chapter 2 is precisely a logic of this sort.

CHAPTER 2

Incorporating the Principle of Reversibility in Stalnaker's Logic

The discussion on counterfactuals, from 1970s onwards, has been primarily devoted to the works established by David Lewis and Robert Stalnaker.¹ Their theories of counterfactuals rely on the possible-world semantics and distance measurements among possible worlds. Despite the emergence of new theories,² Lewis's and Stalnaker's works have retained a wide-acceptance in philosophy. More recently, the structure-based framework of counterfactuals, derived from Judea Pearl's causal modeling, has been gaining ground in philosophical circles.³ In this chapter, an overview of the semantics and logics of both frameworks will be presented. This will be followed by a comparison of Pearl's logic to Stalnaker's logic, which aims to explore the claim that Stalnaker's logic is more analogous to Pearl's than Lewis's in light of the Pearl's principle of definiteness.⁴ Next, a study of Pearl's principle of reversibility will be presented in order to support a suggested translation of this principle in terms of the language of the Stalnaker-Lewis framework. Finally, as the translated principle is not valid in Stalnaker's semantics, an extension of Stalnaker's logic incorporating the translated reversibility will be presented, and it will be shown that this extended system is sound and complete with respect to a modified Stalnaker's semantics.

2.1. Similarity-based Theories of Counterfactuals

In the study of counterfactuals, David Lewis, indeed, is one of the most crucial proponents for the adoption of possible-world semantics.⁵ Lewis has different formulations of the truth condition of a counterfactual conditional, and one of these from Lewis (1973c) is addressed as follows.

¹See Lewis (1973b), Stalnaker (1968), and Stalnaker and Thomason (1970).

²See Bennett (1984), Jackson (1977), and Nute (1975, 1976).

³An notable example is Woodward (2003).

⁴Pearl has offered a comparison between his logic with Lewis's logic. See Galles and Pearl (1998) and Pearl (2009).

⁵For those who demand an investigation on the metaphysics of possible-world semantics, see Lewis (1986b) for an in-depth discussion.

Lewis’s truth-condition of counterfactual conditionals

$A \Box \rightarrow C$ is true at [world] i iff C holds at every closest (accessible) A -world to i , if any (1973c, p. 422).⁶

The “ $\Box \rightarrow$ ” is a non-truth-functional connective denoting the counterfactual conditional. Thus, $A \Box \rightarrow C$ is read as “if it were the case that A , C would be the case”. A ϕ -world is a world in which ϕ is true. Similar to modal logic, whether a world is accessible from a particular world depends on the *accessibility* relation among possible worlds.⁷ Moreover, whether a world is *closer* to a particular world than other worlds is determined by Lewis’s famous notion of *comparative similarity*. We will revisit this notion following the explication of Lewis’s logic of counterfactuals.

Lewis’s favorite formulation of his theory is stated with the help of the *systems of spheres*. However, in view of the comparative work later in this thesis, this formulation is not as convenient as his reformulation in terms of *selection functions*, which is “the simplest and the most direct formulation” (1973b, p. 57). Hereafter, Lewis’s formal semantics in terms of this reformulation will be adopted in this thesis.⁸ A Lewisian model M is a tuple $\langle W, I, f \rangle$,⁹ where W is a non-empty *set of possible worlds*, I is an *interpretation function* which maps each possible world in W to a truth assignment such that the truth of a proposition P in a world w is denoted as $I_w(P) = T$, and f is a *selection function* over W which outputs a set of worlds after inputting a proposition and a possible world. By following a similar labeling from Stalnaker (1968), the inputted possible world, the inputted proposition, and the worlds in the outputted set are named the *base world*, the *antecedent*, and the *selected world(s)* respectively.¹⁰ The selection function purports to pick out the set of the closest antecedent-worlds (the selected worlds which have the antecedent true) relative

⁶This truth-condition is not Lewis’s “final analysis” among his discussed analyses in Lewis (1973c). This version embodied the so-called limit assumption which assumes that there is always a set of closest antecedent-worlds for any world i , given that the antecedent is not impossible. It is not necessarily so, as Lewis argued, since it is possible to have closer and closer antecedent-worlds without an end, and thus there is no such a set of closest antecedent-worlds. See Lewis (1973b, pp. 19-20), and Lewis (1973c, pp. 424-425) for his final analysis. Nevertheless, given my primary goal is to compare Pearl’s logic which characterizes causal models which are finite models, forgoing the limit assumption will cause much inconvenience. Hence, I will take a version of Lewis’s semantics with limit assumption in this thesis for a more analogous comparison with Pearl’s framework, unless otherwise specified.

⁷Lewis (1973b, p. 78) noticed that the accessibility relation can be defined in terms of selection function.

⁸The reformulation in terms of selection function assumes that there is a set of closest antecedent-worlds given a possible counterfactual antecedent. This is the limit assumption mentioned earlier. See footnote 6.

⁹I follow Halpern (2010) and Zhang (2011) of characterizing a Lewisian model where accessibility is not included in the model-schema. See Lewis (1973c) for an original characterization.

¹⁰See Stalnaker (1968, p. 104). Notice that Stalnaker does not allow multiple selected worlds. This point will be fully discussed in the following.

to a base world. The set of closest worlds in which a formula ϕ is true relative to a world w is selected by $f(\phi, w)$.

Let L be the language formed by starting with any propositional variable and closing off under \supset , \sim , and $\Box \rightarrow$. Other connectives (e.g. \wedge , \vee , and \equiv) are defined as usual. Accordingly, the semantics of a formula in L is given as follows. For any $w \in W$, any proposition P , and any formula $\phi, \psi \in L$:

$$\begin{aligned} (M, w) \models P & \quad \text{iff} \quad I_w(P) = \text{T}; \\ (M, w) \models \sim\phi & \quad \text{iff} \quad I_w(\phi) = \text{F}; \\ (M, w) \models \phi \supset \psi & \quad \text{iff} \quad I_w(\phi) = \text{F} \text{ or } I_w(\psi) = \text{T}; \\ (M, w) \models \phi \Box \rightarrow \psi & \quad \text{iff} \quad I_{w'}(\psi) = \text{T} \text{ for every } w' \in f(\phi, w). \end{aligned}$$

In addition, there are four conditions which should be satisfied by a (Lewisian) selection function. For any $w \in W$, and any formula $\phi, \psi \in L$:

- (S1) If ϕ is true at w , then $f(\phi, w) = \{w\}$.
- (S2) ϕ is true at every world in $f(\phi, w)$.
- (S3) If ψ is true at every world at which ϕ is true, and $f(\phi, w) \neq \emptyset$, then $f(\psi, w) \neq \emptyset$.
- (S4) If ψ is true at every world at which ϕ is true, and ϕ is true in some world in $f(\psi, w)$, then $f(\phi, w)$ contains all and only those worlds in $f(\psi, w)$ at which ϕ is true.¹¹

(S1) requires that a world be always closer to itself than any other world. (S2) demands the truth of the antecedent in the selected world. (S3) requires that if a formula is impossible for a world, then any stronger formula is also impossible for that world. Finally, (S4) is a condition aiming to guarantee a consistent similarity measurement to a base world. In other words, it avoids the following case: world w is closer to the base world than another world w' relative to an antecedent ϕ , but w is more distant to the base world than w' relative to another antecedent ψ . Let \mathbf{M} be the class of models in which, for any $M \in \mathbf{M}$, the selection function f in M fulfills the conditions (S1)-(S4).

Next, the logic of counterfactuals developed by Lewis (1973b, p. 132), namely **VC**, will be investigated. The axiomatization of **VC** is as follows:

Rules of inference

- (R1) Modus Ponens
- (R2) If $(\psi_1 \wedge \dots \wedge \psi_n) \supset \chi$ is a theorem, then $((\phi \Box \rightarrow \psi_1) \wedge \dots \wedge (\phi \Box \rightarrow \psi_n)) \supset (\phi \Box \rightarrow \chi)$ is a theorem.

¹¹The formulation of (S1), (S2), (S3), (S4), and (SS) closely follows the formulation in Zhang (2011). See Lewis (1973b, p. 58) for his original formulation.

Axioms schemata

- (VC0) All instances of truth-functional tautologies
- (VC1) $\phi \Box \rightarrow \phi$
- (VC2) $(\sim \phi \Box \rightarrow \phi) \supset (\psi \Box \rightarrow \phi)$
- (VC3) $(\phi \Box \rightarrow \sim \psi) \vee (((\phi \wedge \psi) \Box \rightarrow \chi) \equiv (\phi \Box \rightarrow (\psi \supset \chi)))$
- (VC4) $(\phi \Box \rightarrow \psi) \supset (\phi \supset \psi)$
- (VC5) $(\phi \wedge \psi) \supset (\phi \Box \rightarrow \psi)$

The validity of (VC1) is grounded by (S2). (VC2) is validated by conditions (S2) and (S3). (VC3) is guaranteed by (S4), and (S1) validates both (VC4) and (VC5). Beyond Lewis's work, another important logic of counterfactuals (based on the similarity-based framework) is Stalnaker's work in Stalnaker (1968) and Stalnaker and Thomason (1970). The difference of Stalnaker's logic from Lewis's **VC** is that Stalnaker requires that, given that the antecedent is not impossible to the base world, there is *one and only one* world in the set of closest possible worlds outputted by the selection function. This condition can be captured by an additional condition of selection function (SS).

- (SS) For any formula $\phi \in L$, and any world $w \in W$, $f(\phi, w)$ is a singleton set or \emptyset (i.e. an empty set).¹²

This condition correspondingly validates an axiom (S), which is the famous *law of conditional excluded middle*.

- (S) $(\phi \Box \rightarrow \psi) \vee (\phi \Box \rightarrow \sim \psi)$

Let \mathbf{M}_S be a subclass of \mathbf{M} in which, for any $M \in \mathbf{M}_S$, the selection function f in M fulfills the condition (SS), in addition to (S1)-(S4). Supplementing (S) to **VC** then yields the Stalnaker's logic in which Lewis denotes as **VCS** (Lewis, 1973b, p. 133).

(SS), indeed, is a condition that Lewis does not concede. To claim that world w' is closer to world w than world w'' , in Lewis's theory, is to claim that w' is *more similar* to w than w'' . If the restriction of *unique* closest world is imposed on the selection function, it implies that there cannot be a *tie* in weighting comparative similarity, that is, two or more worlds are outputted by the selection function. As Lewis put it, "(his truth condition) is the obvious revision of Stalnaker's analysis to permit a tie in comparative similarity between several equally close closest A -worlds" (1973c, p. 422). The difference between Lewis's and Stalnaker's can be made more explicit by an example due to Quine (1950).

¹²Note that Stalnaker (1968) employs world-selection function rather than set-selection function, where an absurd world λ is included in the model schema playing the role of the empty set.

It is not the case that if Bizet and Verdi were compatriots, Bizet would be Italian; and it is not the case that if Bizet and Verdi were compatriots, Bizet would not be Italian; nevertheless, if Bizet and Verdi were compatriots, Bizet either would or would not be Italian (1950, p. 14).

By symbolizing “Bizet and Verdi are compatriots” as P and “Bizet is Italian” as Q , we have $\sim(P \Box \rightarrow Q) \wedge \sim(P \Box \rightarrow \sim Q) \wedge (P \Box \rightarrow (Q \vee \sim Q))$. The third conjunct is valid in both Lewis’s and Stalnaker’s semantics. It is the first two conjuncts which are at stake. In Lewis’s aforementioned semantics, particularly with regard to his analysis of overall similarity, he allows the case of two (or more) possible worlds being closest to a base world. Given that Bizet and Verdi are not compatriots in the actual world, it could be the case that they are compatriots in both w_1 and w_2 which are both closest to the actual world, where Bizet is an Italian in w_1 but a French in w_2 . Hence, it is neither the case that Bizet is an Italian in all worlds in the set of the closest compatriot-with-Verdi-worlds, nor the case that Bizet is not an Italian in all these worlds. Nevertheless, Stalnaker’s logic leaves no room of truth for the stated case. Stalnaker does not allow the multiplicity of closest worlds according to his condition (SS), and the first two conjuncts together immediately contradict the axiom (S). Hence, the formula is contradictory in Stalnaker’s logic.¹³

In summary, the distinction between Lewis’s framework and Stalnaker’s rests on the limitation of number of worlds in the set of the closest possible worlds. Given that the set is non-empty, Lewis allows multiple worlds in the set but Stalnaker allows at most one world. This distinction, as will be argued very soon, shows that Stalnaker’s logic is more analogous to Pearl’s structure-based logic than Lewis’s. The following section offers an overview of Pearl’s framework.

2.2. Pearl’s Structure-based Theory of Counterfactuals

Unlike the similarity-based theories, Pearl’s theory offers a *causal* interpretation of counterfactuals based on a *modifiable structural equation model*. Putting it differently, articulating counterfactuals with *causal models*. Roughly speaking, the truth of a counterfactual conditional depends on whether the consequent is true under a *hypothetical intervention* that makes the antecedent

¹³I am not arguing that Stalnaker has imposed a wrong condition on selection function and thus given assurance to an axiom which should never receive credentials. Whether **VC** or **VCS** is more favorable is not at issue for the current purpose of reviewing them. For those who are interested, see Stalnaker (1980) in defending the law of conditional excluded middle.

true.¹⁴ This section will introduce the framework developed by Galles and Pearl (1998) for the later comparative work.¹⁵

A *signature* S for causal models is a triple $\langle \mathbf{U}, \mathbf{V}, R \rangle$ ¹⁶ where \mathbf{U} and \mathbf{V} are finite sets of variables, and R restricts the range of values, which is a finite set, taken by each variable $X \in \mathbf{U} \cup \mathbf{V}$. A causal model T is a tuple $\langle S, F \rangle$ over a signature S where F is a collection of functions, such that for each variable $X \in \mathbf{V}$, there is a unique function $f_X: \times_{Y \in \mathbf{U} \cup \mathbf{V} \setminus \{X\}} R(Y) \rightarrow R(X)$. In other words, the function maps each value configuration of $\mathbf{U} \cup \mathbf{V} \setminus \{X\}$ to the value of X . Thus, the value of $X \in \mathbf{V}$ depends on the values of other variables *inside* the causal models where f_X characterizes the *causal mechanism* of X . Note that for all $U \in \mathbf{U}$, there is no function for U since its value is determined *outside* the model. Therefore, a value configuration of \mathbf{U} plays the role of describing the *context* or *background conditions* for the causal model. Naturally, \mathbf{U} are called the set of *exogenous variables* whereas \mathbf{V} are the set of *endogenous variables*.

A causal model is also known as a *structural equation model* in the sense that each f_X corresponds to a structural equation: $X = f_X(\mathbf{U} \cup \mathbf{V} \setminus \{X\})$. Usually, for every variable $X \in \mathbf{V}$, some variables in $\mathbf{U} \cup \mathbf{V} \setminus \{X\}$ are redundant in determining the value of X . They can then be omitted from the structural equation for X and the set of remaining variables which contributes in determining the value of X are denoted as PA_X (i.e. the parents of X). $G(T)$ is a directed graph which represents the qualitative features of a causal model T and it consists of nodes and directed edges in which the former corresponds to variables in $\mathbf{U} \cup \mathbf{V}$ and the latter points from members of PA_X toward X .

Given a causal model $T = \langle \mathbf{U}, \mathbf{V}, R, F \rangle$, a (possibly empty) subset of endogenous variables $\mathbf{X} \subseteq \mathbf{V}$, and a possible value configuration \mathbf{x} of \mathbf{X} (i.e. \mathbf{x} contains one and only one value for each variable in \mathbf{X}), a submodel $T_{\mathbf{X}=\mathbf{x}}$ denotes the causal model that results from T by replacing the structural equation of \mathbf{X} in T with $\mathbf{X} = \mathbf{x}$. So, excluding that for each $X \in \mathbf{X}$ that the equation for X is modified into $X = x$ (where x is the component value in \mathbf{x} for \mathbf{X}), $T_{\mathbf{X}=\mathbf{x}}$ is the same as the original model T . The modeling which allows the replacement of structural equations thus gains its name *modifiable* structural equation models.

One of the common themes shared by different theories of counterfactuals is how to capture the idea of *minimal change*, or how a hypothetical situation

¹⁴Not every theory of causal modeling interprets counterfactual conditionals with hypothetical intervention. For instance, see Handfield et al. (2007) and Hiddleston (2005).

¹⁵In order to offer a more elegant and rigorous formalism, the modeling method offered by Halpern (2000), Halpern (2010), and Zhang (2011) is adopted in this thesis.

¹⁶Bold letters are used to denote a (possibly empty) set.

$$U \longrightarrow X \longrightarrow Y$$

FIGURE 2.2.1

differs minimally from the actuality. In Pearl’s logic, the functions in F symbolize *independent physical mechanisms*. To change minimally is to modify the function(s) in question, for instance, replacing the functions of T with $\mathbf{X} = \mathbf{x}$ into $T_{\mathbf{X}=\mathbf{x}}$. Intuitively, $T_{\mathbf{X}=\mathbf{x}}$ models the counterfactual situation in which \mathbf{X} is intervened to take the value \mathbf{x} while the mechanisms for other endogenous variables are kept intact.

Next, we are going to consider different classes of causal models. A *solution* to a causal model, relative to a value configuration of \mathbf{u} of the exogenous variables \mathbf{U} , is a value configuration \mathbf{v} of the endogenous variables \mathbf{V} such that all structural equations in the model are simultaneously satisfied. Generally speaking, there may or may not be a solution to a causal model relative to a value configuration of the exogenous variables, and even when there are, there may be more than one solution.

A model T having its signature $S = \langle \mathbf{U}, \mathbf{V}, R \rangle$ is said to have the property of *unique solution* if and only if for every $\mathbf{X} \subseteq \mathbf{V}$, every value configuration \mathbf{x} of \mathbf{X} , and every value configuration \mathbf{u} of \mathbf{U} , there is one and only one solution to the model $T_{\mathbf{X}=\mathbf{x}}$ relative to \mathbf{u} . This class of models which has this property is referred as $\mathbf{T}_{uniq}(S)$. Moreover, $\mathbf{T}_{rec}(S)$, a subclass of $\mathbf{T}_{uniq}(S)$, is the class of models where there is a *total ordering* \rightsquigarrow over all variables in \mathbf{V} , such that if $X \rightsquigarrow Y$, then $Y \notin PA_X$. If $T \in \mathbf{T}_{rec}(S)$, then $T \in \mathbf{T}_{uniq}(S)$. It is because of the obvious fact that one can solve the variables in the order given by \rightsquigarrow . Galles and Pearl (1998) refer this class of models as *recursive models*, or *acyclic models*. If a model is recursive, then there is no *feedback relation* in the model. Finally, $\mathbf{T}_{all}(S)$ refers to the class of all causal models.¹⁷ Here are some examples concerning these different classes:

EXAMPLE 2.2.1. A recursive model

$$T = \langle \mathbf{U}, \mathbf{V}, R, F \rangle$$

$$\mathbf{U} = \{U\}, \mathbf{V} = \{X, Y\}$$

$$R(U) = R(X) = R(Y) = \{0, 1\}$$

$$X = U$$

$$Y = X$$

See $G(T)$ in Figure 2.2.1.

EXAMPLE 2.2.2. A non-recursive model with unique solution

$$T = \langle \mathbf{U}, \mathbf{V}, R, F \rangle$$

¹⁷Zhang (2011) has also considered a class of causal models which is a subclass of $\mathbf{T}_{all}(S)$ but a superclass of $\mathbf{T}_{uniq}(S)$, which is shown to validate certain *Lewisian* properties of counterfactuals. See Zhang (2011) for details.

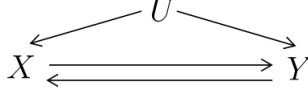


FIGURE 2.2.2

$\mathbf{U} = \{U\}, \mathbf{V} = \{X, Y\}$
 $R(U) = R(X) = R(Y) = \{0, 1\}$
 $X = U \wedge Y$
 $Y = U \vee X$
 See $G(T)$ in Figure 2.2.2.

EXAMPLE 2.2.3. A non-recursive model with multiple solutions
 $T = \langle \mathbf{U}, \mathbf{V}, R, F \rangle$
 $\mathbf{U} = \{U\}, \mathbf{V} = \{X, Y\}$
 $R(U) = R(X) = R(Y) = \{0, 1\}$
 $X = U \vee Y$
 $Y = U \vee X$
 See $G(T)$ in Figure 2.2.2.

The model in Example 2.2.1 is a model in $\mathbf{T}_{rec}(S)$ due to the total ordering over the variables in \mathbf{V} . Given that the exogenous variable U is set, the value of the endogenous variables, X and Y , can be solved one by one by the ordering. When $U = 0$, the unique solution to the model is $\{U = 0, X = 0, Y = 0\}$; similarly, when $U = 1$, the unique solution is $\{U = 1, X = 1, Y = 1\}$. Moreover, in every submodel, there is also a unique solution. For instance, the unique solution to the submodel $T_{X=1}$ is $\{U = 1, X = 1, Y = 1\}$ when $U = 1$, and $\{U = 0, X = 1, Y = 1\}$ is that of the unique solution when $U = 0$. On the other hand, both the models in Example 2.2.2 and Example 2.2.3 are non-recursive models or cyclic models (i.e. both models are not in $\mathbf{T}_{rec}(S)$). As indicated by the two oppositely directed edges between X and Y in their respective directed graph, a feedback relation holds between X and Y in both examples. It is remarkable that the model in Example 2.2.2 is in $\mathbf{T}_{uniq}(S)$ whereas Example 2.2.3 is not (despite their identical qualitative characteristics shown in the directed graphs). In Example 2.2.2, for any value of U and every submodel (including the original model T), there is a unique solution. However, in Example 2.2.3, when $U = 0$, T has two solutions: $\{U = 0, X = 0, Y = 0\}$ and $\{U = 0, X = 1, Y = 1\}$.

In addition, a *basic counterfactual formula* is of the form $[X_1 = x_1 \wedge \dots \wedge X_k = x_k] \phi$, where X_1, \dots, X_k are *distinct variables* in \mathbf{V} , and ϕ is a Boolean combinations of formulas of the form $Y(\mathbf{u}) = y$.¹⁸ It expresses

¹⁸This formulation is provided from Halpern (2000) and the notation of $[\]$ is obviously borrowed from dynamic logic (e.g. Harel 1979). Moreover, it is questionable of how to

the statement “if X_1 were intervened to take value x_1 , ..., and X_k were intervened to take value x_k , it would be the case that ϕ , relative to \mathbf{u} ”. This formula is typically abbreviated as $[\mathbf{X} = \mathbf{x}] \phi$ (i.e. “if the variables in \mathbf{X} were intervened to take the value configuration \mathbf{x} , then ϕ would be the case”) for the sake of convenience. In the special case when \mathbf{X} is empty, the formula $[\mathbf{X} = \mathbf{x}] \phi$ is then written as $[true] \phi$, or simply ϕ . The language which contains all the Boolean combinations of the basic counterfactual formulas is hereafter referred as $L_{CC}(S)$ where “*CC*” stands for *causal counterfactuals*.¹⁹

Given any causal model T over S , every formula in $L_{CC}(S)$ has a truth value. By following Halpern (2000), for a basic counterfactual formula $[\mathbf{X} = \mathbf{x}] \phi$, $T \models [\mathbf{X} = \mathbf{x}] (Y(\mathbf{u}) = y)$ if Y takes the value y in *all* solutions to $T_{\mathbf{X}=\mathbf{x}}$ relative to \mathbf{u} .²⁰ Next, the truth value of any arbitrary counterfactual formulas ϕ and ψ (which is a Boolean combination of basic counterfactual formulas) is defined in the usual way:

$$\begin{aligned} T \models \phi \wedge \psi & \text{ iff } T \models \phi \text{ and } T \models \psi \\ T \models \sim \phi & \text{ iff } T \not\models \phi \end{aligned}$$

A formula ϕ is valid with respect to a class of causal models \mathbf{T} iff $T \models \phi$ for all $T \in \mathbf{T}$.

An example from Pearl (2009, p. 209) helps in illustrating how causal modeling operates. Consider the following model T which describes a case of an execution of a prisoner which takes places in a firing squad:

EXAMPLE 2.2.4. Execution of a prisoner

$$T = \langle \mathbf{U}, \mathbf{V}, R, F \rangle$$

$$\mathbf{U} = \{U\}, \mathbf{V} = \{W, X, Y, Z\}$$

$$R(U) = R(W) = R(X) = R(Y) = R(Z) = \{0, 1\}$$

$$X = U$$

$$Y = X$$

$$Z = X$$

$$W = Y \vee Z$$

See $G(T)$ in Figure 2.2.3.

U denotes whether the court orders the execution of the prisoner; X denotes whether the Captain on the squad gives a signal; Y denotes whether rifleman 1 shoots; Z denotes whether rifleman 2 shoots; and W denotes whether the

handle antecedents in disjunctive form. For a relevant discussion in the Stalnaker-Lewis framework, see Ellis et al. (1977), Lewis (1977), and Loewer (1976).

¹⁹The name of the language is borrowed from Zhang (2011). Halpern (2000) has introduced some other languages for causal modeling when comparing his semantics to the one developed by Galles and Pearl (1998).

²⁰In Halpern (2000), a dual notation $\langle \mathbf{X} = \mathbf{x} \rangle \phi$ is introduced and is defined as $\sim[\mathbf{X} = \mathbf{x}] \sim \phi$. Thus, $T \models \langle \mathbf{X} = \mathbf{x} \rangle (Y(\mathbf{u}) = y)$ if Y takes the value y in *some* solutions to the submodel $T_{\mathbf{X}=\mathbf{x}}$ relative to \mathbf{u} . Note that $T \models [\mathbf{X} = \mathbf{x}] \phi$ is tantamount to $T \models \langle \mathbf{X} = \mathbf{x} \rangle \phi$ when $T \in \mathbf{T}_{\text{uniq}}(S)$ or $T \in \mathbf{T}_{\text{rec}}(S)$, but not when $T \in \mathbf{T}_{\text{all}}(S)$.

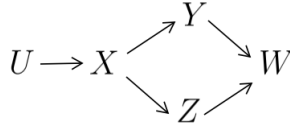


FIGURE 2.2.3

prisoner dies. Let the values 0 and 1 represent non-occurrence and occurrence respectively. Suppose that the court orders the execution, that is, $U = 1$.

Consider the conditional: had the Captain not given a signal, the prisoner would not have died. To express formally, the conditional becomes $[X = 0] (W(1) = 0)$. To determine its truth value, we consider the model $T_{X=0}$ relative to $U = 1$ where every structural equation is the same as T except the equation for X is replaced as $X = 0$.²¹ The (unique) solution to the submodel is $\{U = 1, X = 0, Y = 0, Z = 0, W = 0\}$ and it indicates that W takes value 0 in this solution. Hence, the conditional is true in this model (i.e. $T \models [X = 0] (W(1) = 0)$).

Consider another conditional: had rifleman 1 not shot, the prisoner would not have died, formally speaking, $[Y = 0] (W(1) = 0)$. To evaluate this conditional, we consider the model $T_{Y=0}$ relative to $U = 1$ where every structural equation is the same as T except the equation for Y is replaced as $Y = 0$. The (unique) solution to the submodel is $\{U = 1, X = 1, Y = 0, Z = 1, W = 1\}$ and it indicates that W takes value 1 rather than 0 in this solution. Hence, the conditional is false in this model (i.e. $T \not\models [Y = 0] (W(1) = 0)$).

Next, the following examines different axiomatic systems provided by Halpern (2000) which categorize different classes of causal models:

Rules of inference

(MP) Modus Ponens

Axioms

(C0) All instances of truth-functional tautologies

(C1) $[\mathbf{Y} = \mathbf{y}] (X(\mathbf{u}) = x) \supset [\mathbf{Y} = \mathbf{y}] (X(\mathbf{u}) \neq x')$
if $x, x' \in R(X), x \neq x'$

(Equality)²²

(C2) $\bigvee_{x \in R(X)} [\mathbf{Y} = \mathbf{y}] (X(\mathbf{u}) = x)$

(Definiteness)

(C3) $([\mathbf{X} = \mathbf{x}] (W(\mathbf{u}) = w) \wedge [\mathbf{X} = \mathbf{x}] (Y(\mathbf{u}) = y))$
 $\supset [\mathbf{X} = \mathbf{x} \wedge W = w] (Y(\mathbf{u}) = y)$

(Composition)

²¹ $U = 1$ is essential in evaluating the conditional. As Pearl puts, “the counterfactual consequent must be evaluated under the same background conditions as those prevailing in the actual world” (2009, p. 211).

²²Originally, this axiom is named *uniqueness* in Galles and Pearl (1998).

$$(C4) \quad [\mathbf{X}=\mathbf{x} \wedge W=w] (W(\mathbf{u})=w) \quad (\text{Effectiveness})$$

$$(C5) \quad ([\mathbf{X}=\mathbf{x} \wedge W=w] (Y(\mathbf{u})=y) \wedge [\mathbf{X}=\mathbf{x} \wedge Y=y] (W(\mathbf{u})=w)) \\ \supset [\mathbf{X}=\mathbf{x}] (Y(\mathbf{u})=y), \text{ if } Y \neq W \quad (\text{Reversibility})$$

(C1) states that there is *at most one* solution to a model. (C2), on the other hand, states that there is *at least one* solution to a model. (C3) expresses the idea that if the value of W is w in every solution to $T_{\mathbf{X}=\mathbf{x}}$, then all solutions to $T_{\mathbf{X}=\mathbf{x}, W=w}$ are the same as the solutions to $T_{\mathbf{X}=\mathbf{x}}$. (C4) is simply that an intervention is effective in the sense that W has the value assigned by the intervention. Finally, (C5) says that, if Y takes the value y by forcing W to the value w , and W takes the value w by forcing Y to the value y , then W and Y will have the values w and y respectively without any intervention.

Beyond these five axioms, an extra axiom is needed to characterize the class of recursive models. Particularly, the notation \rightsquigarrow is helpful in axiomatizing the logic for $T_{rec}(S)$ where $Y \rightsquigarrow Z$, read “ Y affects Z ”, is an abbreviation for the formula:

$$\bigvee_{\mathbf{X} \subseteq \mathbf{V}, x \in_{\mathbf{x}} X \in \mathbf{V} R(X), y_a \in R(Y), \mathbf{u} \in_{\mathbf{u}} \times U \in \mathbf{U} R(U), z_c \neq z_d \in R(Z)} \\ ([\mathbf{X}=\mathbf{x} \wedge Y=y_a](Z(\mathbf{u})=z_c) \wedge [\mathbf{X}=\mathbf{x}](Z(\mathbf{u})=z_d)) \quad (\text{Affect})$$

In plain words, this formula expresses the idea that the changed value of Y alters the value of Z under certain setting of some variables, both exogenous and endogenous. The extra axiom is then:

$$(C6) \quad ((X_0 \rightsquigarrow X_1) \wedge \dots \wedge (X_{k-1} \rightsquigarrow X_k)) \\ \supset \sim (X_k \rightsquigarrow X_0) \quad (\text{Recursiveness})$$

Let $\mathbf{AX}_{uniq}(S)$ consist of (C0)-(C5) and MP, and let $\mathbf{AX}_{rec}(S)$ consist of (C0)-(C4), (C6), and MP.²³ Halpern (2000) showed that $\mathbf{AX}_{uniq}(S)$ ($\mathbf{AX}_{rec}(S)$) is a sound and complete axiomatization for $L_{CC}(S)$ with respect to $\mathbf{T}_{uniq}(S)$ ($\mathbf{T}_{rec}(S)$).²⁴

Indeed, many of the axioms are quite plausible intuitively, but reversibility arguable is not. As Halpern (2000) indicated, the validity of reversibility in $\mathbf{T}_{uniq}(S)$ is “non-obvious” (2000, p. 329). Neither is it clear how reversibility

²³(C5) is not included in $\mathbf{AX}_{rec}(S)$ since it can be deduced from (C3) and (C6). See Galles and Pearl (1998) and Halpern (2000). In addition, see Halpern (2000) for a sound and complete axiomatization to $\mathbf{T}_{all}(S)$.

²⁴ $\mathbf{AX}_{uniq}(S)$ and $\mathbf{AX}_{rec}(S)$ are actually not the sets of axioms Pearl used to characterize the class of models with unique solutions and recursive models respectively. However, as pointed out by Halpern (2000), the language employed by Pearl’s is inadequate in expressing certain fundamental axioms like (C1) and (C2) (pp.324-325). Thus, this thesis takes Halpern’s result as a refinement of Pearl’s logic and refers Pearl’s logic as this refined work.

can relate the mentioned distinction between Lewis's and Stalnaker's logics. These are the issues that will be addressed in the following section.

2.3. Reversibility and the Problem of Translation

Reversibility is not as intuitive as the other axioms in $\mathbf{AX}_{uniq}(S)$. I now reproduce the soundness proof of (C5) in $\mathbf{T}_{uniq}(S)$, which is provided by Galles and Pearl (1998) and Halpern (2000).

THEOREM 2.3.1. [Galles and Pearl, Halpern] (C5) *is valid in* $\mathbf{T}_{uniq}(S)$.

PROOF. Let $T \in \mathbf{T}_{uniq}(S)$. Suppose that $T \models [\mathbf{X}=\mathbf{x} \wedge W=w]$ $(Y(\mathbf{u}) = y) \wedge T \models [\mathbf{X}=\mathbf{x} \wedge Y=y] (W(\mathbf{u}) = w)$. I aim to show that $T \models [\mathbf{X}=\mathbf{x}] (Y(\mathbf{u}) = y)$. Given that T is in $\mathbf{T}_{uniq}(S)$, every submodel of T has a unique solution. Let \mathbf{v}_1 be the unique solution to the submodel $T_{\mathbf{X}=\mathbf{x}, W=w}$ (relative to \mathbf{u}) and \mathbf{v}_2 be the unique solution to the submodel $T_{\mathbf{X}=\mathbf{x}, Y=y}$ (relative to \mathbf{u}). Consider the submodel $T_{\mathbf{X}=\mathbf{x}, W=w, Y=y}$ (relative to \mathbf{u}). Note that for any variable $Z \notin \mathbf{X} \cup \{W, Y\}$, the equations for Z in $T_{\mathbf{X}=\mathbf{x}, W=w}$, $T_{\mathbf{X}=\mathbf{x}, Y=y}$, and $T_{\mathbf{X}=\mathbf{x}, W=w, Y=y}$ are all the same, except that W is set to w in $T_{\mathbf{X}=\mathbf{x}, W=w}$, Y is set to y in $T_{\mathbf{X}=\mathbf{x}, Y=y}$, and both W and Y are set to w and y respectively in $T_{\mathbf{X}=\mathbf{x}, W=w, Y=y}$. However, given that w and y are the values of W and Y respectively in both \mathbf{v}_1 and \mathbf{v}_2 , and both of these solutions satisfy the equations of Z in $T_{\mathbf{X}=\mathbf{x}, W=w}$ and $T_{\mathbf{X}=\mathbf{x}, Y=y}$, thus \mathbf{v}_1 and \mathbf{v}_2 are both solutions to the submodel $T_{\mathbf{X}=\mathbf{x}, W=w, Y=y}$. Again, since every submodel of T has a unique solution, $\mathbf{v}_1 = \mathbf{v}_2$. By then I need to show that \mathbf{v}_1 is the unique solution of $T_{\mathbf{X}=\mathbf{x}}$ (relative to \mathbf{u}). As shown above, \mathbf{v}_1 satisfies the equation for Z in $T_{\mathbf{X}=\mathbf{x}}$, for any $Z \notin \mathbf{X} \cup \{W, Y\}$, and since \mathbf{v}_1 satisfies the equation for Y in $T_{\mathbf{X}=\mathbf{x}, W=w}$, \mathbf{v}_1 satisfies the equation for Y in $T_{\mathbf{X}=\mathbf{x}}$ as well. Similarly, since $\mathbf{v}_2 = \mathbf{v}_1$ satisfies the equation for W in $T_{\mathbf{X}=\mathbf{x}, Y=y}$, \mathbf{v}_1 satisfies the equation for W in $T_{\mathbf{X}=\mathbf{x}}$ as well. Given that $T_{\mathbf{X}=\mathbf{x}}$ has a unique solution, finally, as Y has the value y in \mathbf{v}_1 , and \mathbf{v}_1 is the solution to $T_{\mathbf{X}=\mathbf{x}}$, hence $T \models [\mathbf{X}=\mathbf{x}] (Y(\mathbf{u}) = y)$ as desired. \square

By considering reversibility with $\mathbf{X} = \mathbf{x}$ being taken as empty, that is, no intervention is taken other than $W = w$ and $Y = y$ in the antecedent of the principle, a restricted form of reversibility follows:

$$(C5s) \quad ([W=w](Y(\mathbf{u})=y) \wedge [Y=y](W(\mathbf{u})=w)) \supset Y(\mathbf{u})=y, \\ \text{if } Y \neq W \qquad \qquad \qquad \text{(Restricted Reversibility)}$$

A natural way of reading (C5s) is to view the actual state as an equilibrium state attained by the mutual interdependence between two variables, for instance, price and quantity in economics.²⁵ As noticed by both Galles and Pearl (1998) and Halpern (2000), reversibility follows from the property of unique solution.

²⁵This reading is borrowed from Halpern (2000) when he explains the property of unique solution.

It is easy to see that reversibility is invalid in models with multiple solutions, or with no solution. For instance, in [Example 2.2.3](#), even if we have $T \models [X=1](Y(0) = 1)$ and $T \models [Y=1](X(0) = 1)$, it is not the case that $T \models Y(0)=1$.

The property of unique solution is expressed by equality and definiteness, that is, (C1) and (C2). (C1) states that there is *at most one* solution to every submodel, and (C2) states that there is *at least one* solution to every submodel. Thus, they together imply that there is a unique solution to every submodel, which is exactly what the property of unique solution states.

In order to clarify the issues of whether reversibility is valid in the similarity-based theories, a more pressing inquiry is that of the property of unique solution. Nonetheless, such a mission cannot be done easily without a *translation method* bridging up the languages of the two frameworks. The following is the method provided by Galles and Pearl (1998, p.15) for translating a counterfactual conditional expressed in L into $L_{CC}(S)$:

Translation from L into $L_{CC}(S)$

For any sets of variables X and Y , let ϕ and ψ stand for $X = x$ and $Y = y$ respectively, then $[X = x](Y(\mathbf{u}) = y) \equiv \phi \Box \rightarrow \psi$.²⁶

Once the translation is available, the analysis as to whether equality and definiteness are valid in M and M_S can proceed. First, every variable in $\mathbf{U} \cup \mathbf{V}$ will be specified as binary to assimilate the two-valuedness in the similarity-based framework. Let ϕ and ψ stand for $\mathbf{Y} = \mathbf{y}$ and $X = x$ respectively.

$$(C1t) \quad (\phi \Box \rightarrow \psi) \supset \sim(\phi \Box \rightarrow \sim\psi) \tag{Translated Equality}$$

$$(C2t) \quad (\phi \Box \rightarrow \psi) \vee (\phi \Box \rightarrow \sim\psi) \tag{Translated Definiteness}²⁷$$

As Lewis (1973b, pp.79-80) observed, (C1t) is valid in both Lewis's and Stalnaker's semantics, except in a vacuous case that ϕ is an impossible antecedent. Moreover, given that every hypothetical intervention is possible in causal modeling, the vacuous case can then be ignored. However, (C2t), which is exactly the axiom (S) of **VCS**, is valid only in Stalnaker's semantics but not Lewis's.²⁸

This result confirms the previous claim in this thesis that Stalnaker's logic is more analogous to Pearl's than Lewis's in light of the principle of definiteness.

²⁶The value configuration \mathbf{u} of the set of exogenous variables is not expressed in the similarity-based language on the right hand side. But it is not necessary as long as \mathbf{u} is kept constant all the way through. Halpern (2010) has used a much more rigorous but complex way in taking caring of the value configuration \mathbf{u} in the similarity-based language.

²⁷The translated (C2) and (C3) are both shown to be valid in the similarity-based semantics, both Lewis's and Stalnaker's. See Galles and Pearl (1998) and Halpern (2010).

²⁸A similar result is provided by Halpern (2010).

The upcoming enquiry is whether reversibility is valid in Stalnaker's logic. Given that reversibility follows from the property of unique solution, and both equality and definiteness are contained in Stalnaker's framework, the answer seems to be positive. Here is a preliminary translation of reversibility. Let ϕ , ψ , and χ stand for $\mathbf{X} = \mathbf{x}$, $Y = y$ and $W = w$ respectively.

$$(C5t1) \quad (((\phi \wedge \chi) \Box \rightarrow \psi) \wedge ((\phi \wedge \psi) \Box \rightarrow \chi)) \supset (\phi \Box \rightarrow \psi)$$

(Translated Reversibility - Version 1)

This version of translated reversibility has an obvious defect. By substituting ϕ with any arbitrary formula, say P , and substituting both ψ and χ with the same formula, say Q , it becomes $((P \wedge Q) \Box \rightarrow Q) \wedge ((P \wedge Q) \Box \rightarrow Q) \supset (P \Box \rightarrow Q)$. Obviously, Q is always true in the closest $(P \wedge Q)$ -world, but then it follows that $(P \Box \rightarrow Q)$ is true, that is, every counterfactual conditional will be true. Thus, this version can hardly be a principle for this defect. However, substituting both ψ and χ with the same formula is objectionable due to the supplementary condition $Y \neq W$ of (C5). Thus, an appropriate translation should capture the *distinctness* of the two variables. An immediate suggestion is to emphasize the idea that the value of a variable can be *manipulated independently* from the value of another variable. In other words, every value combination of ψ and χ is possible. For the sake of simplicity, a pair of modal operators will be defined to facilitate the latter versions of translation.

DEFINITION 2.3.2. Modal operators²⁹

Let \perp stand for a sentential constant falsehood (e.g. $\phi \wedge \sim \phi$).

For any formula $\phi \in L$,

$$\Diamond \phi \quad =^{df} \quad \sim (\phi \Box \rightarrow \perp)$$

$$\Box \phi \quad =^{df} \quad \sim \phi \Box \rightarrow \perp$$

$\Diamond \phi$ and $\Box \phi$ represent “it is possible that ϕ ” and “it is necessary that ϕ ” respectively. To express the idea that the variables at stake can be independently manipulated, a putative suggested translation is to add as an antecedent $\Diamond(\psi \wedge \chi) \wedge \Diamond(\psi \wedge \sim \chi) \wedge \Diamond(\sim \psi \wedge \chi) \wedge \Diamond(\sim \psi \wedge \sim \chi)$ to (C5t1). It results the following:

$$(C5t2) \quad (\Diamond(\psi \wedge \chi) \wedge \Diamond(\psi \wedge \sim \chi) \wedge \Diamond(\sim \psi \wedge \chi) \wedge \Diamond(\sim \psi \wedge \sim \chi))$$

$$\supset (((\phi \wedge \chi) \Box \rightarrow \psi) \wedge ((\phi \wedge \psi) \Box \rightarrow \chi)) \supset (\phi \Box \rightarrow \psi)$$

(Translated Reversibility - Version 2)

This version evades the absurdity that every counterfactual conditional is true. However, one may still argue that if ψ and χ are contradictory to each other, (C5t2) becomes vacuously true due to the impossibility of $(\psi \wedge \chi)$. A rebuttal to this claim is simple; since distinct variables can be independently manipulated,

²⁹This definition is borrowed from Lewis. See Lewis (1973a, p. 22).

it suggests that they cannot represent identical, overlapping, or even conceptually related events. The case in which ψ and χ are contradictory indeed suggests that ψ and χ stand in a conceptual relation.

But (C5t2) is not immune to revision. A notable case is when ϕ is taken as $(\sim\psi \wedge \sim\chi)$, that is, ϕ is inconsistent with ψ and also with χ . In this case, (C5t2) becomes false even if the antecedent supplemented is satisfied.³⁰ In order to avoid this case, a stronger antecedent is needed in which ϕ has to be involved in the precondition of distinctness.

$$\begin{aligned} \text{(C5t3)} \quad & (\diamond(\phi \wedge \psi \wedge \chi) \wedge \diamond(\phi \wedge \psi \wedge \sim\chi) \wedge \diamond(\phi \wedge \sim\psi \wedge \chi) \wedge \\ & \diamond(\phi \wedge \sim\psi \wedge \sim\chi)) \supset (((\phi \wedge \chi) \Box \rightarrow \psi) \wedge ((\phi \wedge \psi) \Box \rightarrow \chi)) \\ & \supset (\phi \Box \rightarrow \psi) \end{aligned}$$

(Translated Reversibility - Version 3)

(C5t3) is my final analysis of translating (C5) into the language for similarity-based theories. Unlike (C5t1), ψ and χ cannot be the same formula, and unlike (C5t2), the possibility that ϕ is inconsistent with ψ and χ does not falsify (C5t3). Nevertheless, this version is not valid in \mathbf{M}_S .

EXAMPLE 2.3.3. Counter-model to (C5t3) in \mathbf{M}_S ³¹

$$M = \langle W, I, f \rangle \in \mathbf{M}_S$$

$$W = \{w_1, w_2, w_3, w_4, w_5\}$$

$$I_{w_1}(P) = \text{T}, I_{w_1}(Q) = \text{F}, I_{w_1}(R) = \text{F},$$

$$I_{w_2}(P) = \text{T}, I_{w_2}(Q) = \text{T}, I_{w_2}(R) = \text{T},$$

$$I_{w_3}(P) = \text{T}, I_{w_3}(Q) = \text{F}, I_{w_3}(R) = \text{T},$$

$$I_{w_4}(P) = \text{T}, I_{w_4}(Q) = \text{T}, I_{w_4}(R) = \text{F};$$

$$f(P, w_1) = \{w_1\}, f(P \wedge Q, w_1) = \{w_2\}, f(P \wedge R, w_1) = \{w_2\}.$$

Hence, after the trial of different translations on distinctness of variables, reversibility is not valid in Stalnaker's semantics. Either there is a more refined version of reversibility which is eventually valid in Stalnaker's semantics, or there is no better translation and reversibility always has counter-models. The former option does not seem workable. Probably, it is the special feature of structural equation models which contributes to the validity of the principle of reversibility under the property of unique solution. In other words, the property of unique closest world in Stalnaker's semantics is weaker than the property of unique solution in causal modeling. It is interesting, then, to consider extending Stalnaker's logic by embedding the principle of reversibility.

³⁰As ϕ is inconsistent with χ , $((\phi \wedge \chi) \Box \rightarrow \psi)$ is vacuously true. Similarly, the inconsistency between ϕ and ψ results the vacuous truth of $((\phi \wedge \psi) \Box \rightarrow \chi)$ and also the falsity of $(\phi \Box \rightarrow \psi)$. Hence, even if the antecedent supplemented is satisfied, (C5t2) is false in this case.

³¹I have not specified the selection function for the base worlds other than w_1 and also formulas other than those at stake. These details are not essential for the current purpose of falsifying (C5t3) in Stalnaker's semantics.

2.4. Incorporating Reversibility in Stalnaker's Logic - VCSR

In this section, a soundness and completeness proof of a new axiomatic system, which is resulted from **VCS** and (C5t3), to a strengthened Stalnaker's semantics will be provided (with respect to L). Let **VCSR** be the axiomatic system supplementing (C5t3) to **VCS**. Before exploring the strengthened semantics, a definition of *distinctness of formulas* will help in capturing the distinctness of variables discussed in the last section.

DEFINITION 2.4.1. Distinctness of formulas

Relative to a model $M = \langle W, I, f \rangle$, any two formulas $\psi, \chi \in L$ are *distinct* relative to a formula $\phi \in L$ in $w \in W$ iff $I_w(\diamond(\phi \wedge \psi \wedge \chi)) = I_w(\diamond(\phi \wedge \psi \wedge \sim\chi)) = I_w(\diamond(\phi \wedge \sim\psi \wedge \chi)) = I_w(\diamond(\phi \wedge \sim\psi \wedge \sim\chi)) = \mathbf{T}$.

Next, in order to validate (C5t3), an additional condition will be supplemented into the selection function f .

(SR) For any formula $\phi, \psi, \chi \in L$ and any world $w, w', w'', w''' \in W$, if ψ and χ are distinct relative to ϕ in w , then if $I_{w'}(\chi) = I_{w''}(\psi) = \mathbf{T}$ for every $w' \in f(\phi \wedge \psi, w)$ and $w'' \in f(\phi \wedge \chi, w)$, then $I_{w'''}(\psi) = \mathbf{T}$ for every $w''' \in f(\phi, w)$.

Let \mathbf{M}_{SR} be a subclass of \mathbf{M}_S in which, for any $M \in \mathbf{M}_{SR}$, the selection function f in M fulfills the conditions (S1)-(S4), (SS) and (SR).

THEOREM 2.4.2. **VCSR** is a sound axiomatization for L with respect to \mathbf{M}_{SR} .

PROOF. Lewis (1973a) has proven that **VCS** is a sound axiomatization for L with respect to \mathbf{M}_S . Thus, it suffices to provide the validity of (C5t3) in \mathbf{M}_{SR} . For the sake of contradiction, suppose that (C5t3) is not valid in \mathbf{M}_{SR} . It means that there is a model $M \in \mathbf{M}_{SR}$ where $M = \langle W, I, f \rangle$, in which there is a world $w \in W$ such that $\diamond(\phi \wedge \psi \wedge \chi) \wedge \diamond(\phi \wedge \psi \wedge \sim\chi) \wedge \diamond(\phi \wedge \sim\psi \wedge \chi) \wedge \diamond(\phi \wedge \sim\psi \wedge \sim\chi)$ and $((\phi \wedge \chi) \Box \rightarrow \psi) \wedge ((\phi \wedge \psi) \Box \rightarrow \chi)$ are true in w , but $(\phi \Box \rightarrow \psi)$ is false in w . According to Definition 2.4.1, the former entails that ψ and χ are distinct relative to ϕ in w . It then implies that $f(\phi \wedge \psi \wedge \chi, w), f(\phi \wedge \psi \wedge \sim\chi, w), f(\phi \wedge \sim\psi \wedge \chi, w), f(\phi \wedge \sim\psi \wedge \sim\chi, w)$ are non-empty sets, so are $f(\phi \wedge \psi, w), f(\phi \wedge \chi, w)$, and $f(\phi, w)$. From (SS), they are all singleton sets. From the truth of $((\phi \wedge \chi) \Box \rightarrow \psi) \wedge ((\phi \wedge \psi) \Box \rightarrow \chi)$ in w , $I_{w'}(\psi) = \mathbf{T}$ and $I_{w''}(\chi) = \mathbf{T}$ for any $w' \in f(\phi \wedge \psi, w)$ and $w'' \in f(\phi \wedge \chi, w)$. Similarly, $I_{w'''}(\psi) = \mathbf{F}$ for any $w''' \in f(\phi, w)$ follows from the falsity of $(\phi \Box \rightarrow \psi)$ in w . However, a contradiction arises with $I_{w'''}(\psi) = \mathbf{T}$ which follows from (SR) and the above. Hence, there is no such model M and (C5t3) is therefore valid in \mathbf{M}_{SR} . \square

Next, the completeness proof will be demonstrated by following the standard strategy of employing *canonical models*.³²

DEFINITION 2.4.3. **VCSR**-consistency

A set of formulas Γ in L is said to be **VCSR**-consistent, or simply consistent, iff there is no $\{\phi_1, \dots, \phi_n\} \subseteq \Gamma$ such that $\vdash_{\mathbf{VCSR}} \sim(\phi_1 \wedge \dots \wedge \phi_n)$.

DEFINITION 2.4.4. Maximal consistency

A set of formulas Γ in L is said to be maximally consistent iff Γ is consistent and for any formula $\phi \in L$, either $\phi \in \Gamma$ or $\sim\phi \in \Gamma$.

LEMMA 2.4.5. *Let Γ be any maximally consistent set of formulas in L . The following holds for any formula $\phi, \psi \in L$:*

- 2.4.5.1 $\phi \in \Gamma$ iff $\sim\phi \notin \Gamma$.
- 2.4.5.2 $\phi \vee \psi \in \Gamma$ iff either $\phi \in \Gamma$ or $\psi \in \Gamma$.
- 2.4.5.3 $\phi \wedge \psi \in \Gamma$ iff $\phi \in \Gamma$ and $\psi \in \Gamma$.
- 2.4.5.4 If $\vdash_{\mathbf{VCSR}} \phi$, then $\phi \in \Gamma$.
- 2.4.5.5 If $\phi \in \Gamma$ and $\vdash_{\mathbf{VCSR}} \phi \supset \psi$, then $\psi \in \Gamma$.
- 2.4.5.6 If $\phi \in \Gamma$ and $\phi \supset \psi \in \Gamma$, then $\psi \in \Gamma$.
- 2.4.5.7 If $\phi \in \Gamma$ and $\phi \equiv \psi \in \Gamma$, then $\psi \in \Gamma$.
- 2.4.5.8 If $\sim\phi \in \Gamma$ and $\phi \equiv \psi \in \Gamma$, then $\sim\psi \in \Gamma$.

PROOF. Obvious from the definition of maximal consistency. See Hughes and Cresswell (1996, p. 114). \square

LEMMA 2.4.6. *Lindenbaum's Lemma*

Every consistent set of formulas in L can be extended to a maximally consistent set.

PROOF. By the standard construction. See Hughes and Cresswell (1996, pp. 115-116). \square

DEFINITION 2.4.7. Canonical model

The canonical model M is a tuple $\langle W, I, f \rangle$ such that

- 2.4.7.1 $W = \{w : w \text{ is a maximally consistent set of formulas in } L\}$;
i.e. W is the set containing all and only maximally consistent sets of formulas in L .
- 2.4.7.2 For any $w \in W$ and any propositional variable P ,
 $I_w(P) = \text{T}$ iff $P \in w$.
- 2.4.7.3 For any $w \in W$ and any formula $\phi \in L$,
 $S(\phi, w) = \{\psi : (\phi \Box \rightarrow \psi) \in w\}$.

³²For a textbook demonstration of this strategy, see e.g. Hughes and Cresswell (1996, ch. 6).

2.4.7.4 For any $w \in W$ and any formula $\phi \in L$,

$$f(\phi, w) = \begin{cases} \{S(\phi, w)\} & \text{if } S(\phi, w) \text{ is consistent,} \\ \emptyset & \text{otherwise.} \end{cases}$$

To prove that f is indeed a selection function, we need the following lemma.

LEMMA 2.4.8. *$S(\phi, w)$ is either a maximally consistent set of formulas or a set containing every formula in L .*

PROOF. $S(\phi, w)$ is a set of formulas which is either consistent or inconsistent. Suppose that it is consistent. Since w is a maximally consistent set according to Definition 2.4.7.1, from the axiom schema (S) (i.e. $\vdash_{\mathbf{VCSR}} (\phi \Box \rightarrow \psi) \vee (\phi \Box \rightarrow \sim \psi)$), Lemma 2.4.5.4 and Lemma 2.4.5.2, either $(\phi \Box \rightarrow \psi) \in w$ or $(\phi \Box \rightarrow \sim \psi) \in w$, for any formula $\phi, \psi \in L$. By Definition 2.4.7.3, either $\psi \in S(\phi, w)$ or $\sim \psi \in S(\phi, w)$. Therefore, $S(\phi, w)$ is maximal according to Definition 2.4.4 and thus maximally consistent. On the other hand, suppose that $S(\phi, w)$ is an inconsistent set of formulas. From Definition 2.4.3, there is a set of formulas $\{\psi_1, \dots, \psi_n\} \subseteq S(\phi, w)$ such that $\vdash_{\mathbf{VCSR}} \sim(\psi_1 \wedge \dots \wedge \psi_n)$. It entails that for any $\chi \in L$, $\vdash_{\mathbf{VCSR}} (\psi_1 \wedge \dots \wedge \psi_n) \supset \chi$. It follows from (R2) that $\vdash_{\mathbf{VCSR}} ((\phi \Box \rightarrow \psi_1) \wedge \dots \wedge (\phi \Box \rightarrow \psi_n)) \supset (\phi \Box \rightarrow \chi)$. From $\{\psi_1, \dots, \psi_n\} \subseteq S(\phi, w)$, it follows from Definition 2.4.7.3 and Lemma 2.4.5.3 that $((\phi \Box \rightarrow \psi_1) \wedge \dots \wedge (\phi \Box \rightarrow \psi_n)) \in w$. It then entails from Lemma 2.4.5.6 that $(\phi \Box \rightarrow \chi) \in w$. From Definition 2.4.7.3, it entails that $\chi \in S(\phi, w)$ for any $\chi \in L$. Hence $S(\phi, w)$ is either a maximally consistent set of formulas or an inconsistent set of formulas which contains every formula in L . \square

LEMMA 2.4.9. *In the canonical model, for any $w \in W$ and any formula $\phi \in L$, $f(\phi, w)$ is either a subset of W or \emptyset .*

PROOF. The lemma easily follows from Lemma 2.4.8 and Definition 2.4.7.4. \square

The next lemma aims to show that, according to the semantical rules in Stalnaker's semantics, a formula is true in a world in the canonical model if and only if the formula belongs to that world.

LEMMA 2.4.10. *In the canonical model, for any $w \in W$ and any formula $\phi \in L$, $I_w(\phi) = \mathbf{T}$ iff $\phi \in w$.*

PROOF. We prove the lemma by structural induction. The base case follows from Definition 2.4.7.2. For the inductive step, we prove the following:

- (a) If the lemma holds for ϕ , it also holds for $\sim \phi$.
- (b) If the lemma holds for ϕ and ψ , it also holds for $\phi \supset \psi$.

(c) If the lemma holds for ϕ and ψ , it also holds for $\phi \Box \rightarrow \psi$.

PROOF (a): Consider a formula $\sim\phi \in L$ and any $w \in W$. $I_w(\sim\phi) = T$ iff $I_w(\phi) = F$. Suppose that the lemma holds for ϕ . $I_w(\phi) = F$ iff $\phi \notin w$. It follows from Definition 2.4.4 that $I_w(\sim\phi) = T$ iff $\sim\phi \in w$.

PROOF (b): Consider a formula $(\phi \supset \psi) \in L$ and any $w \in W$. $I_w(\phi \supset \psi) = T$ iff $I_w(\phi) = F$ or $I_w(\psi) = T$. Suppose that the lemma holds for ϕ and ψ . $I_w(\phi \supset \psi) = T$ iff $\phi \notin w$ or $\psi \in w$. From Definition 2.4.4 and $\phi \notin w$, it follows that $\sim\phi \in w$. Thus, $I_w(\phi \supset \psi) = T$ iff $\sim\phi \in w$ or $\psi \in w$. From Lemma 2.4.5.2, $I_w(\phi \supset \psi) = T$ iff $\sim\phi \vee \psi \in w$. Hence, by the standard definition of \supset , $I_w(\phi \supset \psi) = T$ iff $\phi \supset \psi \in w$.

PROOF (c): Consider a formula $(\phi \Box \rightarrow \psi) \in L$ and any $w \in W$. $I_w(\phi \Box \rightarrow \psi) = T$ iff $I_{w'}(\psi) = T$ for any $w' \in f(\phi, w)$. Suppose that the lemma holds for ϕ and ψ .

[Only if] Suppose that $I_w(\phi \Box \rightarrow \psi) = T$. It follows that $I_{w'}(\psi) = T$ for any $w' \in f(\phi, w)$. Since that lemma holds for ψ , it entails that $\psi \in w'$ for any $w' \in f(\phi, w)$. Further suppose that $S(\phi, w)$ is consistent. From Definition 2.4.7.4, it entails that $\psi \in S(\phi, w)$. It follows from Definition 2.4.7.3 that $(\phi \Box \rightarrow \psi) \in w$. Suppose that $S(\phi, w)$ is inconsistent. From Definition 2.4.3, there is a set of formulas $\{\chi_1, \dots, \chi_n\} \subseteq S(\phi, w)$ such that $\vdash_{\mathbf{VCSR}} \sim(\chi_1 \wedge \dots \wedge \chi_n)$. It implies that $\vdash_{\mathbf{VCSR}} (\chi_1 \wedge \dots \wedge \chi_n) \supset \psi$. From (R2), it follows that $\vdash_{\mathbf{VCSR}} ((\phi \Box \rightarrow \chi_1) \wedge \dots \wedge (\phi \Box \rightarrow \chi_n)) \supset (\phi \Box \rightarrow \psi)$. From $\{\chi_1, \dots, \chi_n\} \subseteq S(\phi, w)$, it follows from Definition 2.4.7.3 and Lemma 2.4.5.3 that $((\phi \Box \rightarrow \chi_1) \wedge \dots \wedge (\phi \Box \rightarrow \chi_n)) \in w$. By Lemma 2.4.5.5, $(\phi \Box \rightarrow \psi) \in w$.

[If] Suppose that $(\phi \Box \rightarrow \psi) \in w$. It follows from Definition 2.4.7.3 that $\psi \in S(\phi, w)$. Suppose that $S(\phi, w)$ is consistent. From Definition 2.4.7.4, it entails that $\psi \in w'$ for any $w' \in f(\phi, w)$. Since the lemma holds for ψ , it follows that $I_{w'}(\psi) = T$ for any $w' \in f(\phi, w)$, and thus $I_w(\phi \Box \rightarrow \psi) = T$. If $S(\phi, w)$ is inconsistent, it follows from Definition 2.4.7.4 that $f(\phi, w) = \emptyset$. Then $I_{w'}(\psi) = T$ for any $w' \in f(\phi, w)$ follows trivially, and thus $I_w(\phi \Box \rightarrow \psi) = T$. \square

Next, I prove that the f in the canonical model satisfies the constraints on the selection function, i.e. (S1)-(S4), (SS) and (SR). I show them one by one.

LEMMA 2.4.11. *f in the canonical model fulfills (S1)*

For every $w \in W$ and any formula $\phi \in L$, if ϕ is true at w , then $f(\phi, w) = \{w\}$.

PROOF. Suppose that ϕ is true at w , that is, $I_w(\phi) = T$. It follows from Lemma 2.4.10 that $\phi \in w$. To start with, I prove the following two claims: (1) for any formula $\psi \in S(\phi, w)$, $\psi \in w$; and (2) for any formula $\psi \in w$, $\psi \in S(\phi, w)$. First, Definition 2.4.7.3, for any $\psi \in S(\phi, w)$, $(\phi \Box \rightarrow \psi) \in w$. From (VC4) (i.e. $\vdash_{\mathbf{VCSR}} ((\phi \Box \rightarrow \psi) \supset (\phi \supset \psi))$), $\phi \in w$, Lemma 2.4.5.5 and Lemma 2.4.5.6,

$\psi \in w$. Hence, (1) is proven. Secondly, for any formula $\psi \in w$, since $\phi \in w$, it follows from Lemma 2.4.5.3 that $(\phi \wedge \psi) \in w$. From (VC5) (i.e. $\vdash_{\mathbf{VCSR}} ((\phi \wedge \psi) \supset (\phi \Box \rightarrow \psi))$), and Lemma 2.4.5.5, $(\phi \Box \rightarrow \psi) \in w$. By Definition 2.4.7.3, it follows that $\psi \in S(\phi, w)$ and thus (2) is proven. From (1) and (2), it entails that $S(\phi, w) = w$. Since w is a maximally consistent set, $S(\phi, w) = w$ is consistent. From Definition 2.4.7.4, it follows that $f(\phi, w) = \{S(\phi, w)\}$. Hence, $f(\phi, w) = \{w\}$. \square

LEMMA 2.4.12. *f in the canonical model fulfills (S2)*

For every $w \in W$ and any formula $\phi \in L$, ϕ is true at every world in $f(\phi, w)$.

PROOF. From (VC1) (i.e. $\vdash_{\mathbf{VCSR}}(\phi \Box \rightarrow \phi)$), by Definition 2.4.7.3 and Lemma 2.4.5.4, it follows that $\phi \in S(\phi, w)$. Suppose that $S(\phi, w)$ is consistent. It then follows from Definition 2.4.7.4 that $f(\phi, w) = \{S(\phi, w)\}$. From Lemma 2.4.10, $I_{w'}(\phi) = \mathbf{T}$ for any $w' \in f(\phi, w)$. Hence, ϕ is true at every world in $f(\phi, w)$. Suppose that $S(\phi, w)$ is inconsistent. It follows from Definition 2.4.7.4 that $f(\phi, w) = \emptyset$. Thus, ϕ is true at every world in $f(\phi, w)$ vacuously. \square

LEMMA 2.4.13. *f in the canonical model fulfills (S3)*

For every $w \in W$ and any formula $\phi, \psi \in L$, if ψ is true at every world at which ϕ is true, and $f(\phi, w) \neq \emptyset$, then $f(\psi, w) \neq \emptyset$.

PROOF. Suppose that ψ is true at every world ϕ is true. It suffices to prove that if $f(\psi, w) = \emptyset$, then $f(\phi, w) = \emptyset$. So, suppose that $f(\psi, w) = \emptyset$. It follows from Definition 2.4.7.4 that $S(\psi, w)$ is inconsistent. From Lemma 2.4.8, $S(\psi, w)$ contains all formulas in L , including $\sim\psi$. It then follows that $(\psi \Box \rightarrow \sim\psi) \in w$ by Definition 2.4.7.3. From (VC2) (i.e. $\vdash_{\mathbf{VCSR}}(\psi \Box \rightarrow \sim\psi) \supset (\phi \Box \rightarrow \sim\psi)$), it entails by that Lemma 2.4.5.5 that $(\phi \Box \rightarrow \sim\psi) \in w$. From Definition 2.4.7.3, $\sim\psi \in S(\phi, w)$. By reductio, suppose that $S(\phi, w)$ is consistent. It follows from Definition 2.4.7.4 that $f(\phi, w) = \{S(\phi, w)\}$. From Lemma 2.4.12, ϕ is true at every world in $f(\phi, w)$. Thus, ϕ is true at $S(\phi, w)$. However, since ψ is true at every world ϕ is true, it follows from Lemma 2.4.10 that $\psi \in S(\phi, w)$. $\psi \in S(\phi, w)$ and $\sim\psi \in S(\phi, w)$ together entail that $S(\phi, w)$ is not maximally consistent. From Lemma 2.4.8, $S(\phi, w)$ is an inconsistent set of formulas. Hence, by Definition 2.4.7.4, $f(\phi, w) = \emptyset$. \square

The proof of (SS) will be provided before that of (S4) for a more convenient proof of the latter.

LEMMA 2.4.14. f in the canonical model fulfills (SS)

For every $w \in W$ and any formula $\phi \in L$, $f(\phi, w)$ is a singleton set or \emptyset .

PROOF. The lemma trivially follows from Definition 2.4.7.3 and Definition 2.4.7.4. \square

LEMMA 2.4.15. f in the canonical model fulfills (S4)

For every $w \in W$ and any formula $\phi, \psi \in L$, if ψ is true at every world ϕ is true, and ϕ is true in some world in $f(\psi, w)$, then $f(\phi, w)$ consists all and only those worlds in $f(\psi, w)$ at which ϕ is true.

PROOF. Suppose that ψ is true at every world ϕ is true. Further suppose that ϕ is true in some world in $f(\psi, w)$. From Lemma 2.4.14, $f(\psi, w) \neq \emptyset$ and $I_{w'}(\phi) = \text{T}$ where $\{w'\} = \{S(\psi, w)\} = f(\psi, w)$. From Lemma 2.4.10 and Definition 2.4.7.3, $\phi \in S(\psi, w)$ and $(\psi \Box \rightarrow \phi) \in w$. Suppose that $(\psi \Box \rightarrow \sim \phi) \in w$. It follows from Definition 2.4.7.3 that $\sim \psi \in S(\psi, w)$ and it violates the fact that $w' = S(\psi, w)$ is maximally consistent. Thus, $(\psi \Box \rightarrow \phi) \notin w$ and so $(\psi \Box \rightarrow \phi) \in w$ by Lemma 2.4.5.1.

It would be useful to derive the following claims first. Here are two instances of (VC3): $\vdash_{\mathbf{VCSR}}(\psi \Box \rightarrow \sim \phi) \vee (((\phi \wedge \psi) \Box \rightarrow \chi) \equiv (\psi \Box \rightarrow (\phi \supset \chi)))$ and $\vdash_{\mathbf{VCSR}}(\psi \Box \rightarrow \sim \phi) \vee (((\phi \wedge \psi) \Box \rightarrow \sim \chi) \equiv (\psi \Box \rightarrow (\phi \supset \sim \chi)))$. From Lemma 2.4.5.4, Lemma 2.4.5.2 and $(\psi \Box \rightarrow \phi) \in w$, we have (i) $((\phi \wedge \psi) \Box \rightarrow \chi) \equiv (\psi \Box \rightarrow (\phi \supset \chi)) \in w$ and (ii) $((\phi \wedge \psi) \Box \rightarrow \sim \chi) \equiv (\psi \Box \rightarrow (\phi \supset \sim \chi)) \in w$.

The fulfillment of (S4) of f can be proven by following four claims: (1) if $f(\phi, w) = \emptyset$, then $f(\psi, w) = \emptyset$; (2) if $f(\psi, w) = \emptyset$, then $f(\phi, w) = \emptyset$; (3) for any formula $\chi \in L$, if $\chi \in S(\phi, w)$, then $\chi \in S(\psi, w)$; and (4) for any formula $\chi \in L$, if $\chi \in S(\psi, w)$, then $\chi \in S(\phi, w)$. Note that (3) and (4) are needed instead of the claim “ $f(\phi, w)$ consists all and only those worlds in $f(\psi, w)$ at which ϕ is true” due to Lemma 2.4.14 and the supposition that ϕ is true in every world in $f(\psi, w)$. (2) has been proven by $f(\psi, w) \neq \emptyset$.

To prove (1), suppose that $f(\phi, w) = \emptyset$. From Definition 2.4.7.4, $S\{\phi, w\}$ is an inconsistent set of formulas which contains every formula in L , including $\sim \psi$. It follows from Lemma 2.4.10 that $I_{w''}(\psi) = \text{F}$ for any $w'' \in f(\phi, w)$. From Lemma 2.4.12, ϕ is true in every world in $f(\phi, w)$. Then, a contradiction arises from the supposition that ψ is true at every world ϕ is true. Hence, $f(\phi, w) \neq \emptyset$ and (1) is proven.

Before proving (3) and (4), by reductio, suppose that $(\phi \Box \rightarrow \sim \psi) \in w$. From Definition 2.4.7.3, $\sim \psi \in S(\phi, w)$. Given that $f(\phi, w) \neq \emptyset$, from Definition 2.4.7.4 and Lemma 2.4.10, $I_{w''}(\psi) = \text{F}$ for any $w'' \in f(\phi, w)$. However, as it follows from Lemma 2.4.12 that ϕ is true in every world in $f(\phi, w)$, a contradiction arises with the initial supposition that ψ is true at every world ϕ

is true. Hence, $(\phi \Box \rightarrow \sim \psi) \notin w$ and so $\sim(\phi \Box \rightarrow \sim \psi) \in w$ by Lemma 2.4.5.1. From an instance of (VC3) that $\vdash_{\mathbf{VCSR}}(\phi \Box \rightarrow \sim \psi) \vee (((\phi \wedge \psi) \Box \rightarrow \chi) \equiv (\phi \Box \rightarrow (\psi \supset \chi)))$, Lemma 2.4.5.4 and Lemma 2.4.5.2, either $(\phi \Box \rightarrow \sim \psi) \in w$ or $(((\phi \wedge \psi) \Box \rightarrow \chi) \equiv (\phi \Box \rightarrow (\psi \supset \chi))) \in w$. From $\sim(\phi \Box \rightarrow \sim \psi) \in w$, it follows that (iii) $(((\phi \wedge \psi) \Box \rightarrow \chi) \equiv (\phi \Box \rightarrow (\psi \supset \chi))) \in w$.

I claim that $f(\psi, w) = f(\phi \wedge \psi, w)$. By reductio, suppose that there is a formula $\chi \in L$ such that either $\chi \in S(\psi, w)$ and $\chi \notin S(\phi \wedge \psi, w)$, or $\chi \in S(\phi \wedge \psi, w)$ and $\chi \notin S(\psi, w)$. First, suppose that $\chi \in S(\psi, w)$ and $\chi \notin S(\phi \wedge \psi, w)$. If $S(\phi \wedge \psi, w)$ is an inconsistent set of formulas, by Lemma 2.4.8, any formulas in L is in $S(\phi \wedge \psi, w)$, including χ which contradicts $\chi \notin S(\phi \wedge \psi, w)$. Thus, $S(\phi \wedge \psi, w)$ is consistent (and thus maximally consistent by Lemma 2.4.8) and $\sim \chi \in S(\phi \wedge \psi, w)$ follows from Lemma 2.4.5.1. It entails from Definition 2.4.7.3 that $((\phi \wedge \psi) \Box \rightarrow \sim \chi) \in w$. From (ii) and Lemma 2.4.5.7, it follows that $(\psi \Box \rightarrow (\phi \supset \sim \chi)) \in w$. By Definition 2.4.7.3, $(\phi \supset \sim \chi) \in S(\psi, w)$. From $\phi \in S(\psi, w)$ and Lemma 2.4.5.6, $\sim \chi \in S(\psi, w)$. Since $S(\psi, w) = w'$ which is maximally consistent, a contradiction arises from $\chi \in S(\psi, w)$ and $\sim \chi \in S(\psi, w)$. Therefore, it entails from the supposition that $\chi \in S(\phi \wedge \psi, w)$ and $\chi \notin S(\psi, w)$. It follows from $S(\psi, w) = w'$ and Lemma 2.4.5.1 that $\sim \chi \in S(\psi, w)$, which further derives that $(\phi \supset \sim \chi) \in S(\psi, w)$. Thus, $(\psi \Box \rightarrow (\phi \supset \sim \chi)) \in w$ follows from Definition 2.4.7.3. From (ii) and Lemma 2.4.5.7, $((\phi \wedge \psi) \Box \rightarrow \sim \chi) \in w$ follows. By Definition 2.4.7.3 again, $\sim \chi \in S(\phi \wedge \psi, w)$ is entailed. Then, $S(\phi \wedge \psi, w)$ is not maximally consistent. By Lemma 2.4.8 and Definition 2.4.7.4, $f(\phi \wedge \psi, w) = \emptyset$ and $I_w((\phi \wedge \psi) \Box \rightarrow \chi) = \text{T}$ follows vacuously. From Lemma 2.4.5.3, $\sim \chi \in S(\psi, w)$ and $\phi \in S(\psi, w)$, $(\phi \wedge \sim \chi) \in S(\psi, w)$, equivalently, $(\phi \supset \chi) \notin S(\psi, w)$. It follows from Definition 2.4.7.3 that $\sim(\psi \Box \rightarrow (\phi \supset \chi)) \in w$. From Lemma 2.4.5.8 and (i), $\sim((\phi \wedge \psi) \Box \rightarrow \chi) \in w$ follows. By Lemma 2.4.8, $I_w((\phi \wedge \psi) \Box \rightarrow \chi) = \text{F}$ and contradiction arises. Hence, $S(\psi, w) = S(\phi \wedge \psi, w) = w'$ and thus $f(\psi, w) = f(\phi \wedge \psi, w)$.

To prove (3), suppose that $\chi \in S(\phi, w)$ for any $\chi \in L$. From $S(\phi, w) \neq \emptyset$ and Definition 2.4.7.4, $I_{w''}(\chi) = \text{T}$ for any $w'' \in f(\phi, w)$. It follows that $I_{w''}(\psi \supset \chi) = \text{T}$ for any $w'' \in f(\phi, w)$. By Lemma 2.4.10 and Definition 2.4.7.3, $(\phi \Box \rightarrow (\psi \supset \chi)) \in w$. From Lemma 2.4.5.7 and (iii), $((\phi \wedge \psi) \Box \rightarrow \chi) \in w$. By Definition 2.4.7.3, $\chi \in S(\phi \wedge \psi, w)$. Given that $S(\psi, w) = S(\phi \wedge \psi, w)$, $\chi \in S(\psi, w)$.

To prove (4), suppose that $\chi \in S(\psi, w)$ for any $\chi \in L$. Given that $S(\psi, w) = S(\phi \wedge \psi, w)$, $\chi \in S(\phi \wedge \psi, w)$. From Definition 2.4.7.3, $((\phi \wedge \psi) \Box \rightarrow \chi) \in w$. From (iii) and Lemma 2.4.5.7, $(\phi \Box \rightarrow (\psi \supset \chi)) \in w$ follows. From Definition 2.4.7.3, $(\psi \supset \chi) \in S(\phi, w)$. Since ψ is true at every world ϕ is true, given that ϕ is true in $S(\phi, w)$ by Lemma 2.4.12, it follows from Lemma 2.4.10 that $\psi \in S(\phi, w)$. Finally, from Lemma 2.4.5.6, we have $\chi \in S(\phi, w)$. \square

LEMMA 2.4.16. *f* in the canonical model fulfills (SR)

For any formula $\phi, \psi, \chi \in L$ and any world $w, w', w'', w''' \in W$, if ψ and χ are distinct relative ϕ in w , then if $I_{w'}(\chi) = I_{w''}(\psi) = \text{T}$ for every $w' \in f(\phi \wedge \psi, w)$ and $w'' \in f(\phi \wedge \chi, w)$, then $I_{w'''}(\psi) = \text{T}$ for every $w''' \in f(\phi, w)$.

PROOF. Suppose that ψ and χ are distinct relative ϕ in w . It follows from Definition 2.4.1 that $I_w(\diamond(\phi \wedge \psi \wedge \chi)) = I_w(\diamond(\phi \wedge \psi \wedge \sim \chi)) = I_w(\diamond(\phi \wedge \sim \psi \wedge \chi)) = I_w(\diamond(\phi \wedge \sim \psi \wedge \sim \chi)) = \text{T}$. By Lemma 2.4.5.3 and Lemma 2.4.10, $(\diamond(\phi \wedge \psi \wedge \chi) \wedge \diamond(\phi \wedge \psi \wedge \sim \chi) \wedge \diamond(\phi \wedge \sim \psi \wedge \chi) \wedge \diamond(\phi \wedge \sim \psi \wedge \sim \chi)) \in w$. Further suppose that if $I_{w'}(\psi) = I_{w''}(\chi) = \text{T}$ for any $w' \in f(\phi \wedge \psi, w)$ and $w'' \in f(\phi \wedge \chi, w)$. From $I_{w''}(\chi) = \text{T}$ for any $w'' \in f(\phi \wedge \chi, w)$, it follows that $I_w((\phi \wedge \chi) \square \rightarrow \chi) = \text{T}$. It then follows from Lemma 2.4.10 that $((\phi \wedge \chi) \square \rightarrow \chi) \in w$. Similarly, $((\phi \wedge \psi) \square \rightarrow \psi) \in w$ follows from $I_{w'}(\psi) = \text{T}$ for any $w' \in f(\phi \wedge \psi, w)$. From (C5t3) and Lemma 2.4.5.5, it follows that $(\phi \square \rightarrow \psi) \in w$. By Definition 2.4.7.3 and Lemma 2.4.10, $I_{w'''}(\psi) = \text{T}$ for any $w''' \in f(\phi, w)$. \square

LEMMA 2.4.17. *The canonical model is a model in \mathbf{M}_{SR} .*

PROOF. This lemma follows from Lemma 2.4.11 - Lemma 2.4.16. \square

THEOREM 2.4.18. **VCSR** is a complete axiomatization for L with respect to \mathbf{M}_{SR} .

PROOF. Suppose that ϕ is true in all $M = \langle W, I, f \rangle \in \mathbf{M}_{SR}$. I show that $\vdash_{\mathbf{VCSR}} \phi$. For reductio, suppose that it is not the case that $\vdash_{\mathbf{VCSR}} \phi$. From Definition 2.4.3, $\sim \phi$ is **VCSR**-consistent. By Definition 2.4.7.1 and Lemma 2.4.6, $\{\sim \phi\}$ can be extended to a maximally consistent set $w \in W$. By Lemma 2.4.10 and Lemma 2.4.17, there is a $M \in \mathbf{M}_{SR}$ such that $\sim \phi$ is true at w . It would then be a counter-model to ϕ , which contradicts the initial supposition that ϕ is true in all $M \in \mathbf{M}_{SR}$. Hence, $\vdash_{\mathbf{VCSR}} \phi$. \square

To sum up this section, it is proposed that (C5t3) is an appropriate translation of reversibility and it is embedded into **VCS** which results an axiomatic system named **VCSR**. This system is shown to be sound and complete with respect to a new class of models \mathbf{M}_{SR} , which has a new constraint (SR) in the selection function. An interesting property of this system will be proven in the next chapter.

Pearl’s Reversibility and the Irreversibility of Counterfactual Dependence

In his seminal paper “Causation”, Lewis made famous the notion of counterfactual dependence and argued that counterfactual dependence between two distinct events is usually not reversible. In a joint work (Zhang et al. 2012), Zhang, de Clercq, and I showed that a special case of Pearl’s reversibility entails, and is entailed by, the irreversibility of counterfactual dependence. We also showed that Pearl’s semantics rules out only mutual counterfactual dependence, but not cyclic dependence in general. In this chapter, I present two main results. One is a generalization of the result in Zhang et al. (2012), in which I show that the full principle of reversibility is equivalent to a sort of irreversibility of a generalized notion of counterfactual dependence. The second is motivated by the peculiar fact that Pearl’s logic rules out mutual counterfactual dependence but not cyclic counterfactual dependence in general. I show that the logic developed in the previous chapter rules out cyclic counterfactual dependence in general. In other words, although Stalnaker’s logic allows mutual as well as cyclic counterfactual dependence, adding a principle that forbids mutual counterfactual dependence (i.e. the principle of reversibility) into the logic suffices to rule out cyclic counterfactual dependence in general.

3.1. Counterfactual Dependence and its Irreversibility

In his seminal work “Causation”, Lewis (1973a) offered a reductive analysis of causal dependence between two distinct families of events in terms of counterfactual dependence. For any event C , let $O(C)$ represents the proposition that is true if and only if C occurs. Let c_0, c_1, \dots, c_n and e_0, e_1, \dots, e_n be distinct (possible) events such that no two of the c ’s and no two of the e ’s are compossible. A family of events $E = \{e_0, e_1, \dots, e_n\}$ counterfactually depends on a family of events $C = \{c_0, c_1, \dots, c_n\}$ iff the counterfactual conditionals: $O(c_0) \Box \rightarrow O(e_0), O(c_1) \Box \rightarrow O(e_1), \dots, O(c_n) \Box \rightarrow O(e_n)$ are all true. For the sake of simplicity, I usually restrict my consideration to families of events with exactly two members in this thesis and call them *binary family of events*, except for the generalizations stated later in section 3.2. The two members in a binary family represents the *occurrence* and *non-occurrence* of an event. For instance, a binary family of events $E = \{e_0, e_1\}$ counterfactually depends upon a binary family of events $C = \{c_0, c_1\}$ iff the counterfactual conditionals

$O(C) \Box \rightarrow O(E)$ and $\sim O(C) \Box \rightarrow \sim O(E)$ are both true, where c_0 represents the *non-occurrence* of C and c_1 represents the *occurrence* of C , in which $\sim O(C)$ and $O(C)$ are propositions denoting the former and the latter respectively, similarly for E .

As Lewis (1973a) argued, a family of events E causally depends on a family of events C if E counterfactually depends on C . Putting it in words, E causally depends on C if E would have occurred had C occurred, and E would not have occurred had C not occurred.

Moreover, unlike nomic dependence, Lewis suggested that counterfactual dependence is irreversible as a matter of commonplace (1973a, p. 564). That is, for any two distinct families of events E_1 and E_2 , if E_2 counterfactually depends on E_1 , it is not the case that E_1 counterfactually depends on E_2 . In other words, there is no mutual counterfactual dependence between E_1 and E_2 .¹

Since the present purpose is to study Pearl’s principle of reversibility in light of the notion of counterfactual dependence, I will not defend or criticize the irreversibility of counterfactual dependence in detail. Suffice it to say that it is an interesting principle with some prima facie plausibility.²

¹There is debate as to whether the irreversibility of counterfactual dependence is really that common, especially when *backtracking* counterfactuals are allowed. It is a kind of counterfactuals in which the consequent temporally precedes the antecedent, which validates implications from effects to causes. In Lewis (1979), he presents a *standard resolution of vagueness* in reading counterfactuals and argues that backtracking counterfactuals should receive a non-standard but *special* treatment. Moreover, he argues that the special would eventually *slip back* to the standard. Jackson (1977) agrees that there are two kinds of counterfactuals but there is nothing special about backtracking. He provides two different truth conditions for foretracking and backtracking counterfactuals respectively. Bennett (1984) partly agrees with Jackson that backtracking is not of a special kind, but he offers a unified account which is applicable to both foretracking and backtracking. This thesis remain neutral in this debate as long as the present purpose is to advocate a comparative study between Lewis’s theory and Pearl’s. See chapter 6 of Hausman (1998) for a more detailed discussion on backtracking counterfactuals.

²Zhang et al. (2012) provided a simple argument to support the prima facie plausibility. As suggested by some theorists (Shoham 1990; Spirtes et al. 2000, p. 20), token causation is anti-symmetric, if one accepts the Lewisian thesis that counterfactual dependence between distinct events is sufficient for causation (e.g. Paul 2009), then one have to accept that counterfactual dependence between distinct events is also anti-symmetric, thus irreversible. Yet, Maudlin (2007) has provided a seemingly counterexample to the irreversibility of counterfactual dependence. He argues that the event of Kennedy’s not being the president in December 1963 shares a mutual counterfactual dependence with the event of his being assassinated in November 1963 (2007, p. 144). However, there is a catch in his example that the two cited events are not really distinct. As formulated by Lewis (1986a), “two events are distinct if they have nothing in common: they are not identical, neither is a proper part of the other, nor do they have any common part” (1986a, p. 212). In Maudlin’s example, one’s being president has presupposed that he has not been assassinated, which in turn suggests his not being assassinated is a proper part of his being president. Hence, the example does not really threaten the plausibility of irreversibility of counterfactual dependence. Section 3.2 provides a more detailed investigation on the precondition of events’ distinctness.

3.2. Pearl's Reversibility and the Irreversibility of Counterfactual Dependence

As introduced in the last chapter, the logical form of Pearl's principle of reversibility is:

$$(C5) \quad ([\mathbf{X}=\mathbf{x} \wedge W=w](Y(\mathbf{u})=y) \wedge [\mathbf{X}=\mathbf{x} \wedge Y=y](W(\mathbf{u})=w)) \\ \supset [\mathbf{X}=\mathbf{x}](Y(\mathbf{u})=y), \text{ if } Y \neq W$$

(Reversibility)

And a restricted version of reversibility (by taking the set of variables in \mathbf{X} as empty) is:

$$(C5s) \quad ([W=w](Y(\mathbf{u})=y) \wedge [Y=y](W(\mathbf{u})=w)) \supset Y(\mathbf{u})=y, \text{ if } Y \neq W$$

(Restricted Reversibility)

As shown in Zhang et al. (2012), despite its name, (C5s), if restricted to binary variables, entails, and is entailed by, the irreversibility of counterfactual dependence between distinct events. Their proof will be recast as follows. To begin with, suppose that Y and Z are distinct binary variables representing events E_Y and E_Z respectively. Let $Y=1$ represent the proposition that E_Y occurs, i.e. $O(E_Y)$, and $Y=0$ represent the proposition that E_Y does not occur, i.e. $\sim O(E_Y)$. Similarly for Z , E_Z , and $O(E_Z)$. Consider the following two instances of reversibility:

$$([Y=1](Z(\mathbf{u})=1) \wedge [Z=1](Y(\mathbf{u})=1)) \supset Y(\mathbf{u})=1, \text{ if } Y \neq Z \\ ([Y=0](Z(\mathbf{u})=0) \wedge [Z=0](Y(\mathbf{u})=0)) \supset Y(\mathbf{u})=0, \text{ if } Y \neq Z$$

By following the suggested translation advocated in chapter 2, and by substituting terms into propositions of event's (non-)occurrence, we have:³

$$(\diamond(O(E_Y) \wedge O(E_Z)) \wedge \diamond(O(E_Y) \wedge \sim O(E_Z))) \\ \wedge \diamond(\sim O(E_Y) \wedge O(E_Z)) \wedge \diamond(\sim O(E_Y) \wedge \sim O(E_Z))) \\ \supset (((O(E_Y) \square \rightarrow O(E_Z)) \wedge (O(E_Z) \square \rightarrow O(E_Y))) \supset O(E_Y))$$

(1)

$$(\diamond(\sim O(E_Y) \wedge \sim O(E_Z)) \wedge \diamond(\sim O(E_Y) \wedge O(E_Z))) \\ \wedge \diamond(O(E_Y) \wedge \sim O(E_Z)) \wedge \diamond(O(E_Y) \wedge O(E_Z))) \\ \supset (((\sim O(E_Y) \square \rightarrow \sim O(E_Z)) \wedge (\sim O(E_Z) \square \rightarrow \sim O(E_Y))) \\ \supset \sim O(E_Y))$$

(2)

From (1) and (2), it follows that:

$$(\diamond(O(E_Y) \wedge O(E_Z)) \wedge \diamond(O(E_Y) \wedge \sim O(E_Z))) \\ \wedge \diamond(\sim O(E_Y) \wedge O(E_Z)) \wedge \diamond(\sim O(E_Y) \wedge \sim O(E_Z)))$$

³ ϕ in (C5t3) is taken as empty to simulate the restricted reversibility. It can also be done by substituting a propositional tautology into ϕ .

$$\begin{aligned}
& \supset (((O(E_Y) \square \rightarrow O(E_Z)) \wedge (O(E_Z) \square \rightarrow O(E_Y))) \\
& \wedge (\sim O(E_Y) \square \rightarrow \sim O(E_Z)) \wedge (\sim O(E_Z) \square \rightarrow \sim O(E_Y))) \supset \perp
\end{aligned} \tag{3}$$

(3) entails that:

$$\begin{aligned}
& (\diamond(O(E_Y) \wedge O(E_Z)) \wedge \diamond(O(E_Y) \wedge \sim O(E_Z))) \\
& \wedge \diamond(\sim O(E_Y) \wedge O(E_Z)) \wedge \diamond(\sim O(E_Y) \wedge \sim O(E_Z))) \\
& \supset (((O(E_Y) \square \rightarrow O(E_Z)) \wedge (\sim O(E_Y) \square \rightarrow \sim O(E_Z))) \\
& \supset \sim((O(E_Z) \square \rightarrow O(E_Y)) \wedge (\sim O(E_Z) \square \rightarrow \sim O(E_Y))))
\end{aligned} \tag{4}$$

One can notice that $((O(E_Y) \square \rightarrow O(E_Z)) \wedge (\sim O(E_Y) \square \rightarrow \sim O(E_Z)))$ is precisely Lewis's definition of counterfactual dependence of E_Y upon E_Z , whereas $((O(E_Z) \square \rightarrow O(E_Y)) \wedge (\sim O(E_Z) \square \rightarrow \sim O(E_Y)))$ is that of the counterfactual dependence of E_Z upon E_Y . The antecedent of (4), i.e. $(\diamond(O(E_Y) \wedge O(E_Z)) \wedge \diamond(O(E_Y) \wedge \sim O(E_Z)) \wedge \diamond(\sim O(E_Y) \wedge O(E_Z)) \wedge \diamond(\sim O(E_Y) \wedge \sim O(E_Z)))$, highlights that E_Y and E_Z are distinct families of events, that is, each value combination between proposition $O(E_Y)$ and $O(E_Z)$ is possible. So what (4) expresses is precisely the irreversibility of counterfactual dependence: given that E_Y and E_Z are distinct, if E_Y counterfactually depends on E_Z , then E_Z does not counterfactually depend on E_Y .

It is worth mentioning that the antecedent of (4) is indispensable. It means that E_Y and E_Z should not be identical, or overlap or stand somehow in logical or conceptual relations. Thus, as Zhang et al. (2012) noted, (4) only starts sounding plausible if E_Y and E_Z are *suitably distinct* in the sense necessary for them to stand in a causal relation.⁴ Thus, distinctness is stronger than non-identity in this thesis. As explained in chapter 2, the antecedent of (4) captures the condition that Y and Z are distinct variables. It is an implicit convention in causal modeling that distinct variables cannot represent identical, overlapping, or conceptually related events, given that every submodel of a given model is considered to be possible, which means that distinct variables can be independently manipulated. Therefore, distinct variables have to represent suitably distinct families of events, if they represent families of events at all.⁵

The above has already proven that the restricted reversibility, when constrained to binary variables, entails the irreversibility of counterfactual dependence between distinct events. What follows provides a proof for the other

⁴See Paul (2009). And see footnote 2 for Lewis's (1986a) strict definition of distinctness.

⁵As pointed out by de Clercq, the antecedent of (4) can be satisfied by *partially overlapping* events. Thus, the principle of irreversibility states that the two partially overlapping events cannot stand in mutual counterfactual dependence, but in fact they can. Although both Pearl and Lewis would not accept partially overlapping to stand in causal relation, it is not clear how to revise (C5t3) to exclude them.

direction. To make the proof more readable, the distinctness is suppressed for the moment. It follows from (4) that:

$$\begin{aligned} & ((O(E_Y) \Box \rightarrow O(E_Z)) \wedge (O(E_Z) \Box \rightarrow O(E_Y))) \\ & \supset \sim((\sim O(E_Z) \Box \rightarrow \sim O(E_Y)) \wedge (\sim O(E_Z) \Box \rightarrow \sim O(E_Y))) \end{aligned} \quad (5)$$

According to the axiom schema (VC5) (i.e. $(\phi \wedge \psi) \supset (\phi \Box \rightarrow \psi)$) in Lewis's **VC**, which is valid under Pearl's semantics,⁶ the contrapositives of two instances of (VC5) are:

$$\begin{aligned} & \sim(\sim O(E_Z) \Box \rightarrow \sim O(E_Y)) \supset \sim(\sim O(E_Z) \wedge \sim O(E_Y)) \\ & \sim(\sim O(E_Y) \Box \rightarrow \sim O(E_Z)) \supset \sim(\sim O(E_Y) \wedge \sim O(E_Z)) \end{aligned}$$

which entail the following:

$$\begin{aligned} & \sim((\sim O(E_Z) \Box \rightarrow \sim O(E_Y)) \wedge (\sim O(E_Z) \Box \rightarrow \sim O(E_Y))) \\ & \supset (O(E_Z) \vee O(E_Y)) \end{aligned} \quad (6)$$

(5) and (6) together entail that:

$$\begin{aligned} & ((O(E_Y) \Box \rightarrow O(E_Z)) \wedge (O(E_Z) \Box \rightarrow O(E_Y))) \\ & \supset (O(E_Z) \vee O(E_Y)) \end{aligned} \quad (7)$$

By the axiom schema (VC4) (i.e. $(\phi \Box \rightarrow \psi) \supset (\phi \supset \psi)$) in **VC**, which is also valid under Pearl's semantics, an instance of this axiom is:

$$(O(E_Z) \Box \rightarrow O(E_Y)) \supset (O(E_Z) \supset O(E_Y)) \quad (8)$$

By supplementing back the distinctness condition, it is not difficult to see that (1) follows from (7) and (8). The other instances of reversibility (e.g. (2)) can be derived similarly. Therefore, restricted reversibility, when constrained to binary variables, is *precisely* the principle that counterfactual dependence between distinct families of events is irreversible.

I will generalize the above result in three steps. First, I will consider (C5s) where variables that are not necessarily binary. Second, I will consider the more general (C5) with only binary variables. Finally, I will present the most general result, concerning (C5) applied to variables which are not necessarily binary.

First generalization (G1)

Consider two distinct (not necessarily binary) variables Y and Z such that $R(Y) = \{y_0, y_1, \dots, y_m\}$ and $R(Z) = \{z_0, z_1, \dots, z_n\}$ where $m, n \geq 1$. Suppose

⁶Both (VC4) and (VC5) follow from Pearl's principle of composition (i.e. (C3)). See Pearl (2009, pp. 240-241).

that Y and Z respectively correspond to families of events $\{E_{y_0}, E_{y_1}, \dots, E_{y_m}\}$ and $\{E_{z_0}, E_{z_1}, \dots, E_{z_n}\}$, such that $Y = y_i$ (where $y_i \in R(Y)$) represents the proposition that E_{y_i} occurs, i.e. the proposition $O(E_{y_i})$. Similarly for Z , any $z_j \in R(Z)$, E_{z_j} and $O(E_{z_j})$.

DEFINITION 3.2.1. Distinctness of families of events

For any two families of events $E_Y = \{E_{y_0}, E_{y_1}, \dots, E_{y_m}\}$ and $E_Z = \{E_{z_0}, E_{z_1}, \dots, E_{z_n}\}$, E_Y and E_Z are *distinct*, written as $D(E_Y, E_Z)$, iff $\bigwedge_{0 \leq i \leq m, 0 \leq j \leq n} \diamond(O(E_{y_i}) \wedge O(E_{z_j}))$ is true.

This definition helps to simplify the long antecedent of reversibility, especially when the variables are non-binary.

DEFINITION 3.2.2. Pairwise irreversibility of counterfactual dependence

For any two families of events E_Y and E_Z , counterfactual dependence between E_Y and E_Z is *pairwise irreversible*, or say, there is no pairwise mutual counterfactual dependence between them iff the following is true: for any $E_{y_a}, E_{y_b} \in E_Y$ and $E_{z_c}, E_{z_d} \in E_Z$ (where $a \neq b$ and $c \neq d$), if $\{E_{z_c}, E_{z_d}\}$ counterfactually depends on $\{E_{y_a}, E_{y_b}\}$, then it is not the case that $\{E_{y_a}, E_{y_b}\}$ counterfactually depends on $\{E_{z_c}, E_{z_d}\}$.

Irreversibility principle 1 (IP1)

For any two families of events E_Y and E_Z , if E_Y and E_Z are distinct, then counterfactual dependence between them is pairwise irreversible.

I now show that (C5s) is equivalent to the (IP1). First, consider the following instances of restricted reversibility:

$$\begin{aligned} ([Y = y_a] (Z(\mathbf{u}) = z_c) \wedge [Z = z_c] (Y(\mathbf{u}) = y_a)) \supset Y(\mathbf{u}) = y_a, \text{ if } Y \neq Z \\ ([Y = y_b] (Z(\mathbf{u}) = z_d) \wedge [Z = z_d] (Y(\mathbf{u}) = y_b)) \supset Y(\mathbf{u}) = y_b, \text{ if } Y \neq Z \\ \text{where } a \neq b \text{ and } c \neq d \end{aligned}$$

After translation and making use of Definition 3.2.1, we have:

$$\begin{aligned} D(E_Y, E_Z) \supset (((O(E_{y_a}) \square \rightarrow O(E_{z_c})) \\ \wedge (O(E_{z_c}) \square \rightarrow O(E_{y_a}))) \supset O(E_{y_a})) \end{aligned} \tag{1}$$

$$\begin{aligned} D(E_Y, E_Z) \supset (((O(E_{y_b}) \square \rightarrow O(E_{z_d})) \\ \wedge (O(E_{z_d}) \square \rightarrow O(E_{y_b}))) \supset O(E_{y_b})) \end{aligned} \tag{2}$$

(1) and (2) together gives us that:

$$\begin{aligned} D(E_Y, E_Z) \supset (((O(E_{y_a}) \square \rightarrow O(E_{z_c})) \\ \wedge (O(E_{z_c}) \square \rightarrow O(E_{y_a})) \wedge (O(E_{y_b}) \square \rightarrow O(E_{z_d}))) \end{aligned}$$

$$\wedge (O(E_{z_d}) \square \rightarrow O(E_{y_b})) \supset (O(E_{y_a}) \wedge O(E_{y_b})) \quad (3)$$

Since $O(E_{y_a})$ and $O(E_{y_b})$ are not compossible, it follows that:

$$\begin{aligned} D(E_Y, E_Z) &\supset (((O(E_{y_a}) \square \rightarrow O(E_{z_c})) \\ &\wedge (O(E_{z_c}) \square \rightarrow O(E_{y_a})) \wedge (O(E_{y_b}) \square \rightarrow O(E_{z_d})) \\ &\wedge (O(E_{z_d}) \square \rightarrow O(E_{y_b}))) \supset \perp) \end{aligned} \quad (4)$$

By simple rearrangement, (4) entails that:

$$\begin{aligned} D(E_Y, E_Z) &\supset (((O(E_{y_a}) \square \rightarrow O(E_{z_c})) \\ &\wedge (O(E_{y_b}) \square \rightarrow O(E_{z_d}))) \supset \sim((O(E_{z_c}) \square \rightarrow O(E_{y_a}))) \\ &\wedge (O(E_{z_d}) \square \rightarrow O(E_{y_b}))) \end{aligned} \quad (5)$$

(5) expresses exactly what is stated in the (IP1). That is, given that y_a and y_b (also z_c and z_d) are arbitrary values of Y (and Z), (5) states that there is no pairwise mutual counterfactual dependence between distinct families of events E_Y and E_Z . Thus, the principle follows from the restricted reversibility. Next, to show the converse, (5) entails that:

$$\begin{aligned} D(E_Y, E_Z) &\supset (((O(E_{y_a}) \square \rightarrow O(E_{z_c})) \\ &\wedge (O(E_{z_c}) \square \rightarrow O(E_{y_a}))) \supset (\sim(O(E_{y_b}) \square \rightarrow O(E_{z_d})) \\ &\vee \sim(O(E_{z_d}) \square \rightarrow O(E_{y_b})))) \end{aligned} \quad (6)$$

From the axiom (VC5) schema (i.e. $(\phi \wedge \psi) \supset (\phi \square \rightarrow \psi)$) in **VC**, two instances of its contrapositive form are:

$$\begin{aligned} \sim(O(E_{z_d}) \square \rightarrow O(E_{y_b})) &\supset \sim(O(E_{z_d}) \wedge O(E_{y_b})) \\ \sim(O(E_{y_b}) \square \rightarrow O(E_{z_d})) &\supset \sim(O(E_{y_b}) \wedge O(E_{z_d})) \end{aligned}$$

These two instances and (6) entail that:

$$\begin{aligned} D(E_Y, E_Z) &\supset (((O(E_{y_a}) \square \rightarrow O(E_{z_c})) \\ &\wedge (O(E_{z_c}) \square \rightarrow O(E_{y_a}))) \supset (\sim O(E_{y_b}) \vee \sim O(E_{z_d}))) \end{aligned} \quad (7)$$

Since E_{y_b} (and E_{z_d}) is an arbitrary event that is not compossible with E_{y_a} (and E_{z_c}), it follows that:

$$\begin{aligned} D(E_Y, E_Z) &\supset (((O(E_{y_a}) \square \rightarrow O(E_{z_c})) \\ &\wedge (O(E_{z_c}) \square \rightarrow O(E_{y_a}))) \supset (O(E_{y_a}) \vee O(E_{z_c}))) \end{aligned} \quad (8)$$

An instance of the axiom (VC4) schema (i.e. $(\phi \square \rightarrow \psi) \supset (\phi \supset \psi)$) in **VC** is:

$$(O(E_{z_c}) \square \rightarrow O(E_{y_a})) \supset (O(E_{z_c}) \supset O(E_{y_a})) \quad (9)$$

Similar to the simple analysis stated, (1) follows from (8) and (9). Other instances can be derived in a very much similar manner. It shows that restricted reversibility follows from the irreversibility principle 1. Hence, (IP1) is equivalent to (C5s).

Second generalization (G2)

Next, I generalize from (C5s) to (C5) with only binary variables. Consider two distinct binary variables Y and Z which correspond to binary families of events E_Y and E_Z respectively, such that $Y = 1$ represents the proposition that E_Y occurs, i.e. $O(E_Y)$, and $Y = 0$ represents the proposition that E_Y does not occur, i.e. $\sim O(E_Y)$. Similarly for Z , E_Z , and $O(E_Z)$. Also consider a set of variables \mathbf{X} and let it corresponds to a family of events $E_{\mathbf{X}}$, such that for any $X \in \mathbf{X}$, $X = x$ represents the proposition $O(E_X)$, i.e. the occurrence of the event E_X .

DEFINITION 3.2.3. Relative distinctness of families of events

For any two families of events $E_Y = \{E_{y_0}, E_{y_1}, \dots, E_{y_m}\}$, $E_Z = \{E_{z_0}, E_{z_1}, \dots, E_{z_n}\}$, and any event E , E_Y and E_Z are *distinct relative to E*, written as $D(E_Y, E_Z|E)$, iff the following is true:

$$\bigwedge_{0 \leq i \leq m, 0 \leq j \leq n} \diamond(O(E) \wedge O(E_{y_i}) \wedge O(E_{z_j})) .$$

DEFINITION 3.2.4. Relative counterfactual dependence

For any event E , any two families of events E_Y and E_Z , any $E_{y_a}, E_{y_b} \in E_Y$ and $E_{z_c}, E_{z_d} \in E_Z$ (where $a \neq b$ and $c \neq d$), $\{E_{z_c}, E_{z_d}\}$ counterfactually depends on $\{E_{y_a}, E_{y_b}\}$ *relative to E* iff $((O(E) \wedge O(E_{y_a})) \square \rightarrow O(E_{z_c})) \wedge ((O(E) \wedge O(E_{y_b})) \square \rightarrow O(E_{z_d}))$ is true.

The notion of counterfactual dependence *relative to an event* in Definition 3.2.4 has a philosophical motivation. Lewis (1973a) defined the concept of *causal chain* to solve problems of *preemption*. A causal chain is a finite sequence of actual events E_1, E_2, \dots, E_n where E_{i+1} causally depends on E_i for all $i \geq 1$. E_i is a cause of E_j iff there is a causal chain from E_i to E_j . For instance, a senior assassin and a junior assassin are on a mission to shoot a dictator. The senior assassin plays as a backup just in case the junior assassin fails to pull his trigger, possibly due to his nervousness. The junior assassin shoots and the dictator dies. The dictator's death neither counterfactually depends on the junior's shot, nor that of the senior's. However, there is a causal chain between the junior's shot and the dictator's death containing intermediary events, like the speeding of the bullet and the hitting of the dictator's body. But there is no such a causal chain linking up the senior's shot and the

death of the dictator. The concept of causal chain is then handy for Lewis to differentiate the preempted potential cause (i.e. the senior's shot) from the preempting actual cause (i.e. the junior's shot).

Making use of causal chains has the price of embracing transitivity. Not every writer agrees with Lewis that the price is worth paying. Hitchcock (2001), Halpern and Pearl (2005), and Woodward (2003) have employed a very similar concept of counterfactual dependence relative to an event (or events) in their own account of actual causation. To test whether a variable has a casual influence on another variable by hypothetical intervention, the value of other variables has to be fixed. As named by Halpern and Pearl (2005), the value of the other variables needed to be held fixed is the *structural contingency*. In these treatments, no assumption of transitivity is needed but preemption can also be neatly explained with the notion of relative counterfactual dependence.

Then, the irreversibility principle is formulated as:

Irreversibility principle 2 (IP2)

For any two binary families of events E_Y , E_Z , and family of events $E_{\mathbf{X}}$, if E_Y and E_Z are distinct relative to an event $E_X \in E_{\mathbf{X}}$, counterfactual dependence between E_Y and E_Z relative to E_X is irreversible.

I now show that (C5), when Y and Z are restricted to binary variables, is equivalent to (IP2). First, consider the following instances of reversibility:

$$\begin{aligned} & ([X = x \wedge Y = 1] (Z(\mathbf{u}) = 1) \wedge [X = x \wedge Z = 1] (Y(\mathbf{u}) = 1)) \\ & \supset [X = x] (Y(\mathbf{u}) = 1), \text{ if } Y \neq Z \\ & ([X = x \wedge Y = 0] (Z(\mathbf{u}) = 0) \wedge [X = x \wedge Z = 0] (Y(\mathbf{u}) = 0)) \\ & \supset [X = x] (Y(\mathbf{u}) = 0), \text{ if } Y \neq Z \end{aligned}$$

After translation, they become:

$$\begin{aligned} D(E_Y, E_Z | E_X) & \supset (((O(E_X) \wedge O(E_Y)) \square \rightarrow O(E_Z)) \\ & \wedge ((O(E_X) \wedge O(E_Z)) \square \rightarrow O(E_Y))) \supset (O(E_X) \square \rightarrow O(E_Y)) \end{aligned} \tag{1}$$

$$\begin{aligned} D(E_Y, E_Z | E_X) & \supset (((O(E_X) \wedge \sim O(E_Y)) \square \rightarrow \sim O(E_Z)) \\ & \wedge ((O(E_X) \wedge \sim O(E_Z)) \square \rightarrow \sim O(E_Y))) \supset (O(E_X) \square \rightarrow \sim O(E_Y)) \end{aligned} \tag{2}$$

(1) and (2) entail that:

$$\begin{aligned} D(E_Y, E_Z | E_X) & \supset (((O(E_X) \wedge O(E_Y)) \square \rightarrow O(E_Z)) \\ & \wedge ((O(E_X) \wedge O(E_Z)) \square \rightarrow O(E_Y)) \\ & \wedge ((O(E_X) \wedge \sim O(E_Y)) \square \rightarrow \sim O(E_Z)) \\ & \wedge ((O(E_X) \wedge \sim O(E_Z)) \square \rightarrow \sim O(E_Y))) \end{aligned}$$

$$\supset ((O(E_X) \Box \rightarrow O(E_Y)) \wedge (O(E_X) \Box \rightarrow \sim O(E_Y))) \quad (3)$$

Here is an instance of the axiom (C1) (i.e. equality) in Pearl's $\mathbf{AX}_{uniq}(S)$:⁷

$$[X = x](Y(\mathbf{u})=0) \supset [X = x](Y(\mathbf{u}) \neq 1), \text{ if } 0, 1 \in R(X).$$

Its translated form is:

$$(O(E_X) \Box \rightarrow \sim O(E_Y)) \supset \sim(O(E_X) \Box \rightarrow O(E_Y)) \quad (4)$$

It follows from (3) and (4) that:

$$\begin{aligned} D(E_Y, E_Z | E_X) &\supset (((O(E_X) \wedge O(E_Y)) \Box \rightarrow O(E_Z)) \\ &\wedge ((O(E_X) \wedge O(E_Z)) \Box \rightarrow O(E_Y)) \\ &\wedge ((O(E_X) \wedge \sim O(E_Y)) \Box \rightarrow \sim O(E_Z)) \\ &\wedge ((O(E_X) \wedge \sim O(E_Z)) \Box \rightarrow \sim O(E_Y))) \supset \perp \end{aligned} \quad (5)$$

It is easy to see that (5) entails that:

$$\begin{aligned} D(E_Y, E_Z | E_X) &\supset (((O(E_X) \wedge O(E_Y)) \Box \rightarrow O(E_Z)) \\ &\wedge ((O(E_X) \wedge \sim O(E_Y)) \Box \rightarrow \sim O(E_Z))) \\ &\supset \sim(((O(E_X) \wedge O(E_Z)) \Box \rightarrow O(E_Y)) \\ &\wedge ((O(E_X) \wedge \sim O(E_Z)) \Box \rightarrow \sim O(E_Y))) \end{aligned} \quad (6)$$

Manifestly, (6) is the claim that counterfactual dependence between E_Y and E_Z relative to E_X is irreversible. Thus, (C5), when restricted to binary variables, entails (IP2). Next, I show the derivation from the other side. It follows from (6) that:

$$\begin{aligned} D(E_Y, E_Z | E_X) &\supset (((O(E_X) \wedge O(E_Y)) \Box \rightarrow O(E_Z)) \\ &\wedge ((O(E_X) \wedge O(E_Z)) \Box \rightarrow O(E_Y))) \\ &\supset (\sim((O(E_X) \wedge \sim O(E_Y)) \Box \rightarrow \sim O(E_Z)) \\ &\vee \sim((O(E_X) \wedge \sim O(E_Z)) \Box \rightarrow \sim O(E_Y))) \end{aligned} \quad (7)$$

Here are some theorem schemata from Lewis's \mathbf{VC} :

$$\begin{aligned} (\mathbf{VC3}) \quad &(\phi \Box \rightarrow \sim \psi) \vee (((\phi \wedge \psi) \Box \rightarrow \chi) \equiv (\phi \Box \rightarrow (\psi \supset \chi)))^8 \\ (\mathbf{VCt1}) \quad &(\sim((\phi \wedge \psi) \Box \rightarrow \chi) \wedge (\phi \Box \rightarrow (\psi \supset \chi))) \supset (\phi \Box \rightarrow \sim \psi) \\ (\mathbf{VCt2}) \quad &(((\phi \wedge \psi) \Box \rightarrow \chi) \supset (\phi \Box \rightarrow (\psi \supset \chi))) \vee (\phi \Box \rightarrow \sim \psi) \\ (\mathbf{VCt3}) \quad &(((\phi \wedge \psi) \Box \rightarrow \chi) \wedge \sim(\phi \Box \rightarrow (\psi \supset \chi))) \supset (\phi \Box \rightarrow (\psi \supset \chi)) \end{aligned}$$

⁷As discussed in chapter 2, the translated form of equality is valid in Lewis's system in the non-vacuous cases.

⁸(VC3) follows from Pearl's principle of composition (i.e. (C3)), equality (i.e. (C1)) and definiteness (i.e. (C2)).

$$(VCt4) \ ((\phi \wedge \psi) \Box \rightarrow \chi) \supset (\phi \Box \rightarrow (\psi \supset \chi))$$

Both (VCt1) and (VCt2) are derived from (VC3), (VCt3) from (VCt2), and (VCt4) from (VCt3). Two instances of (VCt4) are:

$$\begin{aligned} & ((O(E_X) \wedge O(E_Y)) \Box \rightarrow O(E_Z)) \supset (O(E_X) \Box \rightarrow (O(E_Y) \supset O(E_Z))) \\ & ((O(E_X) \wedge O(E_Z)) \Box \rightarrow O(E_Y)) \supset (O(E_X) \Box \rightarrow (O(E_Z) \supset O(E_Y))) \end{aligned}$$

It follows from these two instances that:

$$\begin{aligned} & (((O(E_X) \wedge O(E_Y)) \Box \rightarrow O(E_Z)) \wedge ((O(E_X) \wedge O(E_Z)) \Box \rightarrow O(E_Y))) \\ & \supset (O(E_X) \Box \rightarrow (O(E_Y) \equiv O(E_Z))) \end{aligned} \tag{8}$$

$$\begin{aligned} & (((O(E_X) \wedge O(E_Y)) \Box \rightarrow O(E_Z)) \wedge ((O(E_X) \wedge O(E_Z)) \Box \rightarrow O(E_Y))) \\ & \supset ((O(E_X) \Box \rightarrow (\sim O(E_Y) \supset \sim O(E_Z))) \\ & \wedge (O(E_X) \Box \rightarrow (\sim O(E_Z) \supset \sim O(E_Y)))) \end{aligned} \tag{9}$$

It is not hard to see that (7) and (9) entail the following:

$$\begin{aligned} & D(E_Y, E_Z | E_X) \supset (((O(E_X) \wedge O(E_Y)) \Box \rightarrow O(E_Z)) \\ & \wedge ((O(E_X) \wedge O(E_Z)) \Box \rightarrow O(E_Y))) \\ & \supset ((\sim ((O(E_X) \wedge \sim O(E_Y)) \Box \rightarrow \sim O(E_Z))) \\ & \wedge (O(E_X) \Box \rightarrow (\sim O(E_Y) \supset \sim O(E_Z)))) \\ & \vee (\sim ((O(E_X) \wedge \sim O(E_Z)) \Box \rightarrow \sim O(E_Y))) \\ & \wedge (O(E_X) \Box \rightarrow (\sim O(E_Z) \supset \sim O(E_Y)))) \end{aligned} \tag{10}$$

Two instances of (VCt1) are:

$$\begin{aligned} & (\sim ((O(E_X) \wedge \sim O(E_Y)) \Box \rightarrow \sim O(E_Z))) \\ & \wedge (O(E_X) \Box \rightarrow (\sim O(E_Y) \supset \sim O(E_Z))) \supset (O(E_X) \Box \rightarrow O(E_Y)) \\ & (\sim ((O(E_X) \wedge \sim O(E_Z)) \Box \rightarrow \sim O(E_Y))) \\ & \wedge (O(E_X) \Box \rightarrow (\sim O(E_Z) \supset \sim O(E_Y))) \supset (O(E_X) \Box \rightarrow O(E_Z)) \end{aligned}$$

Together with (10), it follows from these two instances that:

$$\begin{aligned} & D(E_Y, E_Z | E_X) \supset (((O(E_X) \wedge O(E_Y)) \Box \rightarrow O(E_Z)) \\ & \wedge ((O(E_X) \wedge O(E_Z)) \Box \rightarrow O(E_Y))) \\ & \supset ((O(E_X) \Box \rightarrow O(E_Y)) \vee (O(E_X) \Box \rightarrow O(E_Z))) \end{aligned} \tag{11}$$

(11) and (8) together deduce (1), which is an instance of the translated reversibility. Hence, similar to the result shown in the last generalization, when restricted to binary variables, (C5) is equivalent to (IP2).

Final generalization (G3)

Lastly, the result in (G2) will be generalized with regard to variables which are not necessarily binary. Consider two distinct (possibly non-binary) variables Y and Z such that $R(Y) = \{y_0, y_1, \dots, y_m\}$ and $R(Z) = \{z_0, z_1, \dots, z_n\}$ where $m, n \geq 1$. Suppose that Y and Z correspond to families of events $\{E_{y_0}, E_{y_1}, \dots, E_{y_m}\}$ and $\{E_{z_0}, E_{z_1}, \dots, E_{z_n}\}$ respectively such that $Y = y_i$ (where $y_i \in R(Y)$) represents the proposition that E_{y_i} occurs, i.e. the proposition $O(E_{y_i})$. Similarly for Z , any $z_j \in R(Z)$, E_{z_j} and $O(E_{z_j})$. Also consider a set of variables \mathbf{X} and suppose that it corresponds to a family of events $E_{\mathbf{X}}$, such that for any $X \in \mathbf{X}$, $X = x$ represents the proposition $O(E_X)$, i.e. the occurrence of the event E_X .

DEFINITION 3.2.5. Pairwise irreversibility of relative counterfactual dependence

For any families of events E_Y and E_Z and $E_{\mathbf{X}}$, counterfactual dependence between E_Y and E_Z relative to $E_{\mathbf{X}}$ is pairwise irreversible iff the following is true:

for any event $E_X \in E_{\mathbf{X}}$, and any $E_{y_a}, E_{y_b} \in E_Y$ and $E_{z_c}, E_{z_d} \in E_Z$ (where $a \neq b$ and $c \neq d$), if $\{E_{z_c}, E_{z_d}\}$ counterfactually depends on $\{E_{y_a}, E_{y_b}\}$ relative to E_X , then it is not the case that $\{E_{y_a}, E_{y_b}\}$ counterfactually depends on $\{E_{z_c}, E_{z_d}\}$.

Irreversibility principle 3 (IP3)

For any two families of events E_Y and E_Z and $E_{\mathbf{X}}$, if E_Y and E_Z are distinct relative to an event $E_X \in E_{\mathbf{X}}$, then counterfactual dependence between E_Y and E_Z relative to E_X is pairwise irreversible.

First of all, consider the following instances of reversibility:

$$\begin{aligned} & ([X = x \wedge Y = a] (Z(\mathbf{u}) = c) \wedge [X = x \wedge Z = c] (Y(\mathbf{u}) = a)) \\ & \supset [X = x] (Y(\mathbf{u}) = a), \text{ if } Y \neq Z \\ & ([X = x \wedge Y = b] (Z(\mathbf{u}) = d) \wedge [X = x \wedge Z = d] (Y(\mathbf{u}) = d)) \\ & \supset [X = x] (Y(\mathbf{u}) = b), \text{ if } Y \neq Z \end{aligned}$$

where $a \neq b$ and $c \neq d$

After translation, they become:

$$\begin{aligned} D(E_Y, E_Z | E_X) & \supset (((O(E_X) \wedge O(E_{y_a})) \square \rightarrow O(E_{z_c})) \\ & \wedge ((O(E_X) \wedge O(E_{z_c})) \square \rightarrow O(E_{y_a}))) \supset (O(E_X) \square \rightarrow O(E_{y_a}))) \end{aligned} \tag{1}$$

$$\begin{aligned} D(E_Y, E_Z | E_X) & \supset (((O(E_X) \wedge O(E_{y_b})) \square \rightarrow O(E_{z_d})) \\ & \wedge ((O(E_X) \wedge O(E_{z_d})) \square \rightarrow O(E_{y_b}))) \supset (O(E_X) \square \rightarrow O(E_{y_b}))) \end{aligned} \tag{2}$$

(1) and (2) entail that:

$$\begin{aligned}
D(E_Y, E_Z|E_X) &\supset (((O(E_X) \wedge O(E_{y_a})) \square \rightarrow O(E_{z_c})) \\
&\wedge ((O(E_X) \wedge O(E_{z_c})) \square \rightarrow O(E_{y_a})) \\
&\wedge ((O(E_X) \wedge O(E_{y_b})) \square \rightarrow O(E_{z_d})) \\
&\wedge ((O(E_X) \wedge O(E_{z_d})) \square \rightarrow O(E_{y_b}))) \\
&\supset ((O(E_X) \square \rightarrow O(E_{y_a})) \wedge (O(E_X) \square \rightarrow O(E_{y_b})))
\end{aligned} \tag{3}$$

Here is an instance of the axiom (C1) (i.e. equality) in $\mathbf{AX}_{\text{uniq}}(S)$:

$$[X = x] (Y(\mathbf{u}) = y_a) \supset [X = x] (Y(\mathbf{u}) \neq y_b), \text{ if } y_a \neq y_b \in R(X)$$

By simple translation, it becomes:

$$(O(E_X) \square \rightarrow \sim O(E_{y_a})) \supset \sim (O(E_X) \square \rightarrow O(E_{y_b})) \tag{4}$$

It follows from (3) and (4) that:

$$\begin{aligned}
D(E_Y, E_Z|E_X) &\supset (((O(E_X) \wedge O(E_{y_a})) \square \rightarrow O(E_{z_c})) \\
&\wedge ((O(E_X) \wedge O(E_{z_c})) \square \rightarrow O(E_{y_a})) \\
&\wedge ((O(E_X) \wedge O(E_{y_b})) \square \rightarrow O(E_{z_d})) \\
&\wedge ((O(E_X) \wedge O(E_{z_d})) \square \rightarrow O(E_{y_b}))) \supset \perp
\end{aligned} \tag{5}$$

Similar to the above generalizations, (5) entails that:

$$\begin{aligned}
D(E_Y, E_Z|E_X) &\supset (((O(E_X) \wedge O(E_{y_a})) \square \rightarrow O(E_{z_c})) \\
&\wedge ((O(E_X) \wedge O(E_{y_b})) \square \rightarrow O(E_{z_d}))) \\
&\supset \sim (((O(E_X) \wedge O(E_{z_c})) \square \rightarrow O(E_{y_a})) \\
&\wedge ((O(E_X) \wedge O(E_{z_d})) \square \rightarrow O(E_{y_b})))
\end{aligned} \tag{6}$$

It is evident that (6) is the irreversibility principle 3 defined above. Thus, (IP3) follows from (C5). I now show the other direction. It follows from (6) that:

$$\begin{aligned}
D(E_Y, E_Z|E_X) &\supset (((O(E_X) \wedge O(E_{y_a})) \square \rightarrow O(E_{z_c})) \\
&\wedge ((O(E_X) \wedge O(E_{z_c})) \square \rightarrow O(E_{y_a}))) \\
&\supset (\sim ((O(E_X) \wedge O(E_{y_b})) \square \rightarrow O(E_{z_d})) \\
&\vee \sim ((O(E_X) \wedge O(E_{z_d})) \square \rightarrow O(E_{y_b})))
\end{aligned} \tag{7}$$

From the axiom schema (S) (i.e. $(\phi \square \rightarrow \psi) \vee (\phi \square \rightarrow \sim \psi)$) in \mathbf{VCS} , the theorem schema $\sim (\phi \square \rightarrow \psi) \supset (\phi \square \rightarrow \sim \psi)$ follows. Two instances of this theorem schema are:

$$\begin{aligned}
&\sim ((O(E_X) \wedge O(E_{y_b})) \square \rightarrow O(E_{z_d})) \\
&\supset ((O(E_X) \wedge O(E_{y_b})) \square \rightarrow \sim O(E_{z_d}))
\end{aligned}$$

$$\begin{aligned}
& \sim((O(E_X) \wedge O(E_{z_d})) \Box \rightarrow O(E_{y_b})) \\
& \supset ((O(E_X) \wedge O(E_{z_d})) \Box \rightarrow \sim O(E_{y_b}))
\end{aligned}$$

These instances and (7) entail that:

$$\begin{aligned}
D(E_Y, E_Z|E_X) & \supset (((O(E_X) \wedge O(E_{y_a})) \Box \rightarrow O(E_{z_c})) \\
& \wedge ((O(E_X) \wedge O(E_{z_c})) \Box \rightarrow O(E_{y_a}))) \\
& \supset (((O(E_X) \wedge O(E_{y_b})) \Box \rightarrow \sim O(E_{z_d})) \\
& \vee ((O(E_X) \wedge O(E_{z_d})) \Box \rightarrow \sim O(E_{y_b})))
\end{aligned} \tag{8}$$

Two instances of (VCt4) are:

$$\begin{aligned}
& ((O(E_X) \wedge O(E_{y_b})) \Box \rightarrow \sim O(E_{z_d})) \\
& \supset (O(E_X) \Box \rightarrow (O(E_{y_b}) \supset \sim O(E_{z_d}))) \\
& ((O(E_X) \wedge O(E_{z_d})) \Box \rightarrow \sim O(E_{y_b})) \\
& \supset (O(E_X) \Box \rightarrow (O(E_{z_d}) \supset \sim O(E_{y_b})))
\end{aligned}$$

The following is derived from (8) and these two instances:

$$\begin{aligned}
D(E_Y, E_Z|E_X) & \supset (((O(E_X) \wedge O(E_{y_a})) \Box \rightarrow O(E_{z_c})) \\
& \wedge ((O(E_X) \wedge O(E_{z_c})) \Box \rightarrow O(E_{y_a}))) \\
& \supset (O(E_X) \Box \rightarrow (O(E_{y_b}) \supset \sim O(E_{z_d})))
\end{aligned} \tag{9}$$

Another two instances of (VCt4) are:

$$\begin{aligned}
& ((O(E_X) \wedge O(E_{y_a})) \Box \rightarrow O(E_{z_c})) \\
& \supset (O(E_X) \Box \rightarrow (O(E_{y_a}) \supset O(E_{z_c}))) \\
& ((O(E_X) \wedge O(E_{z_c})) \Box \rightarrow O(E_{y_a})) \\
& \supset (O(E_X) \Box \rightarrow (O(E_{z_c}) \supset O(E_{y_a})))
\end{aligned}$$

(9) and these two instances together entail that:

$$\begin{aligned}
D(E_Y, E_Z|E_X) & \supset (((O(E_X) \wedge O(E_{y_a})) \Box \rightarrow O(E_{z_c})) \\
& \wedge ((O(E_X) \wedge O(E_{z_c})) \Box \rightarrow O(E_{y_a}))) \\
& \supset ((O(E_X) \Box \rightarrow (O(E_{y_b}) \supset \sim O(E_{z_d}))) \\
& \wedge (O(E_X) \Box \rightarrow (O(E_{y_a}) \equiv O(E_{z_c}))))
\end{aligned} \tag{10}$$

Since E_{y_b} (and E_{z_b}) is an arbitrary event that is not compossible with and E_{y_a} (and E_{z_c}), it follows that:

$$\begin{aligned}
D(E_Y, E_Z|E_X) & \supset (((O(E_X) \wedge O(E_{y_a})) \Box \rightarrow O(E_{z_c})) \\
& \wedge ((O(E_X) \wedge O(E_{z_c})) \Box \rightarrow O(E_{y_a}))) \\
& \supset (O(E_X) \Box \rightarrow ((O(E_{y_b}) \supset O(E_{z_c})))
\end{aligned}$$

$$\wedge (O(E_{y_a}) \equiv O(E_{z_c})))))) \quad (11)$$

It then entails that:

$$\begin{aligned} D(E_Y, E_Z|E_X) &\supset (((O(E_X) \wedge O(E_{y_a})) \Box \rightarrow O(E_{z_c})) \\ &\wedge ((O(E_X) \wedge O(E_{z_c})) \Box \rightarrow O(E_{y_a}))) \\ &\supset (O(E_X) \Box \rightarrow ((O(E_{y_b}) \supset O(E_{y_a})))))) \end{aligned} \quad (12)$$

Since E_{y_b} is not compossible with E_{y_a} , it follows that:

$$\begin{aligned} D(E_Y, E_Z|E_X) &\supset (((O(E_X) \wedge O(E_{y_a})) \Box \rightarrow O(E_{z_c})) \\ &\wedge ((O(E_X) \wedge O(E_{z_c})) \Box \rightarrow O(E_{y_a}))) \\ &\supset (O(E_X) \Box \rightarrow O(E_{y_a}))) \end{aligned} \quad (13)$$

Once again, (13) is our translated reversibility. With regard to the axiom-schema (S) in **VCS**, reversibility follows from (IP3). Therefore, when generalized to variables which are not necessarily binary, (C5) is equivalent to (IP3).⁹

These generalizations reveal the equivalence between Pearl's principle of reversibility, in its full generality, and the statement that (relative) counterfactual dependence is (pairwise) irreversibility in Pearl's counterfactual logic (or Stalnaker's).

3.3. Cyclic Counterfactual Dependence and a Peculiarity in Pearl's Logic

As argued in Zhang et al. (2012), treating the claim that counterfactual dependence between distinct (families of) events is always irreversible as a logical principle instead of a mere commonplace is not entirely implausible. Here is their argument. Given that some authors seem to think that token causation is anti-symmetric (e.g. Shoham 1990; Spirtes et al., 2000, p.20), if one also accepts the Lewisian thesis that counterfactual dependence between distinct families of events is sufficient for causation, then one has to accept that counterfactual dependence between distinct families of events is also anti-symmetric.

On top of that, when supplemented with another Lewisian thesis that token causation is transitive, the above argument leads to the thesis that there is no

⁹Nevertheless, unlike (G1) and (G2), reversibility cannot be entailed by the irreversibility with Lewis's **VC** alone. Appendix provides a counter-model (in **M**). It is a conjecture that Pearl's reversibility, given a feasible translation, is equivalent to the law of conditional excluded middle (i.e. (S)) plus the irreversibility of counterfactual dependence. This hunch is motivated by two facts. First, as shown in chapter 2, even though Stalnaker's logic contains (S), reversibility is not valid in it. Second, as suggested by this counter-model, reversibility does not follow from irreversibility if (S) is absent. A detailed investigation of this conjecture awaits another occasion to be fully developed.

cyclic counterfactual dependence, as cyclic counterfactual dependence entails cyclic causation, and any cyclic causation implies mutual causation (by the transitivity of causation).

Basically, a cycle of counterfactual dependence is that, for a sequence $n \geq 2$ of distinct families of events, E_1, \dots, E_k , such that E_{i+1} counterfactually depends on E_i for all $1 \leq i \leq k - 1$, and E_1 counterfactually depends on E_k . Mutual counterfactual dependence is merely a special case by restricting cyclic counterfactual dependence to length two. It is, indeed, hard to see what is *metaphysically* special about cyclic counterfactual dependence of length two.

In the argument just stated, the transitivity of causation is not indisputable in the discussion of causation (Hall 2000; Hitchcock 2001).¹⁰ As Zhang et al. (2012) mentioned, the argument can be simplified by strengthening the premise of no mutual causation to the premise of no cyclic causation, without assuming transitivity. The intuition behind our argument is that, even though the premise of cyclic causation is stronger than that of no mutual causation, they stand and fall together metaphysically. Hence, the argument against mutual counterfactual dependence is as strong as the argument against cyclic counterfactual dependence. We argued that, if a theory endorses the principle of no mutual counterfactual dependence without endorsing that of no cyclic counterfactual dependence, it is *peculiar* since it would amount to an unmotivated discrimination against cycles of length two.¹¹

Pearl's semantics, interestingly, has exactly this peculiarity. Although it contains the principle of no mutual counterfactual dependence between two distinct families of events, it does allow three or more distinct families of events to form a cycle of counterfactual dependence. Zhang et al. (2012) provided a very simple Pearlian model that contains a cycle of counterfactual dependence.¹²

EXAMPLE 3.3.1. A Pearlian model with a cycle of counterfactual dependence

$$T = \langle \mathbf{U}, \mathbf{V}, R, F \rangle$$

$$\mathbf{U} = \{\}, \mathbf{V} = \{X_1, X_2, X_3\}$$

$$R(X_1) = R(X_2) = R(X_3) = \{0, 1\}$$

$$X_1 = \sim X_2 \vee X_3$$

¹⁰Hall (2000) provided some examples in which transitivity of causation is incompatible with the claim that counterfactual dependence is sufficient for causation.

¹¹As suggested by one of the examiners, a possible response to our argument is that Pearl's logic is said to be peculiar and unmotivated only because of the tacit yet questionable assumption of causal transitivity. As abovementioned, causal transitivity is not an indisputable assumption. Contrarily, if it were indisputable, the conclusion of my argument would not be a peculiarity of Pearl's logic, but a self-contradiction of it. I concede that what sounds peculiar (or unmotivated) to one may appear to be perfectly innocent to another. But it is still plausible to suggest the peculiarity given that one does not refute causal transitivity wholesale.

¹²An extremely similar model was used in Halpern (2010) for a different purpose.

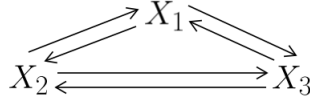


FIGURE 3.3.1

$$X_2 = \sim X_3 \vee X_1$$

$$X_3 = \sim X_1 \vee X_2$$

See $G(T)$ in Figure 3.3.1.

T is in $\mathbf{T}_{uniq}(S)$ since every submodel has a unique solution. In particular, the solution $\{X_1=1, X_2=1, X_3=1\}$ is unique to the submodels $T_{X_1=1}$, $T_{X_2=1}$, and $T_{X_3=1}$. Hence, the counterfactuals $[X_1=1](X_2=1)$, $[X_2=1](X_3=1)$, $[X_3=1](X_1=1)$ are all true in T . More than that, the following three counterfactuals are also true in T : $[X_1=0](X_2=0)$, $[X_2=0](X_3=0)$, $[X_3=0](X_1=0)$, due to the fact that the submodel $T_{X_1=0}$ has the solution $\{X_1=0, X_2=0, X_3=1\}$, $T_{X_2=0}$ has the solution $\{X_1=1, X_2=0, X_3=0\}$, and finally, $T_{X_3=0}$ has the solution $\{X_1=0, X_2=1, X_3=0\}$. By symbolizing X_1, X_2, X_3 , as three distinct families of events E_1, E_2 , and E_3 respectively, the given two sets of counterfactuals imply a cycle of counterfactual dependence of length three: E_2 counterfactually depends on E_1 , E_3 counterfactually depends on E_2 , and E_1 counterfactually depends on E_3 .

This simple model shows that Pearl's logic *merely* rules out cycles of length two. This is indeed the unmotivated discrimination against cycles of length two, as the mentioned argument suggests. Unless there is a justification for this discrimination, it sounds reasonable that Pearl's logic is either too strong by denying the possibility of mutual counterfactual dependence, or too weak by allowing possibility of cyclic counterfactual dependence.¹³

To avoid the peculiarity, on the one hand, Pearl's logic can be weakened to allow any length of cyclic counterfactual dependence, for example, by pursuing an axiomatization in which its corresponding semantics does not contain the property of unique solution. On the other hand, Pearl's logic can be strengthened to deny the possibility of any length of cyclic counterfactual dependence. A possible suggestion is to adopt the logic $AX_{rec}(S)$, which is stronger than $AX_{uniq}(S)$. In the remaining of this chapter, I will argue that adopting $AX_{rec}(S)$ is not necessary, even though it is sufficient, in overcoming the peculiarity. I will suggest that a logic which contains a principle weaker than the principle of recursiveness (i.e. (C6)) can also avoid the peculiarity by strengthening Pearl's logic. First, the following is a principle which aims to

¹³Zhang et al. (2012) have attempted an explanation which might lead to a justification of the peculiarity. See chapter 4 of this thesis for a brief description of this explanation.

tackle the peculiarity by disallowing any length of cyclic counterfactual dependence:

Principle of no cyclic counterfactual dependence (NC)

For any families of events E_1, \dots, E_k , and $E_{\mathbf{X}}$, if E_i and E_j are distinct relative to $E_X \in E_{\mathbf{X}}$ for any i and j ranged from 1 to k (where $i \neq j$), and if $\{E_{i+1_0}, E_{i+1_1}\}$ counterfactually depends on $\{E_{i_0}, E_{i_1}\}$ relative to E_X for all $1 \leq i \leq k-1$, then it is not the case that $\{E_{1_0}, E_{1_1}\}$ counterfactually depends on $\{E_{k_0}, E_{k_1}\}$ relative to E_X .

This principle generalizes the irreversibility of counterfactual dependence, which merely forbids cyclic counterfactual dependence with length two. One can easily spot that (NC) is equivalent to (IP2) by restricting $k = 2$. On the other hand, the following provides a generalized principle of reversibility.

$$\begin{aligned}
(\text{GC5}) \quad & ([\mathbf{X} = \mathbf{x} \wedge Y_1 = y_1] (Y_2(\mathbf{u}) = y_2) \wedge \dots \\
& \wedge [\mathbf{X} = \mathbf{x} \wedge Y_{k-1} = y_{k-1}] (Y_k(\mathbf{u}) = y_k) \\
& \wedge [\mathbf{X} = \mathbf{x} \wedge Y_k = y_k] (Y_1(\mathbf{u}) = y_1)) \supset [\mathbf{X} = \mathbf{x}] (Y_2(\mathbf{u}) = y_2) \\
& \text{where } k \geq 2 \text{ and } Y_i \neq Y_j \text{ for all } i \text{ and } j \text{ within the range from 1 to } k \\
& \text{(Generalized Reversibility)}
\end{aligned}$$

Similarly, (GC5) is exactly (C5) when $k = 2$. Given the shown equivalence between Pearl's reversibility and the irreversibility principle of counterfactual dependence in the last section, it is easy to see (GC5) is equivalent to (NC) when restricted to binary variables.¹⁴

On the other hand, (C6) states that, for any variables X_1, \dots, X_k in \mathbf{V} , if $X_1 \rightsquigarrow X_2, \dots, X_{k-1} \rightsquigarrow X_k$, then it is not the case that $X_k \rightsquigarrow X_1$. Let's recall the definition of $Y \rightsquigarrow Z$ (read: the variable Y affects the variable Z). It is true iff the following is true:

$$\begin{aligned}
& \bigvee_{\mathbf{X} \subseteq \mathbf{V}, x \in_{\times} X \in \mathbf{V} R(X), y_a \in R(Y), \mathbf{u} \in_{\times} U \in \mathbf{U} R(U), z_c \neq z_d \in R(Z)} \\
& ([\mathbf{X} = \mathbf{x} \wedge Y = y_a] (Z(\mathbf{u}) = z_c) \wedge [\mathbf{X} = \mathbf{x}] (Z(\mathbf{u}) = z_d)) \\
& \text{(Affect)}
\end{aligned}$$

For a better comparison with generalized reversibility, it might be useful to show that the following formulation is equivalent to (Affect),

$$\begin{aligned}
& \bigvee_{\mathbf{X} \subseteq \mathbf{V}, x \in_{\times} X \in \mathbf{V} R(X), y_a \neq y_b \in R(Y), \mathbf{u} \in_{\times} U \in \mathbf{U} R(U), z_c \neq z_d \in R(Z)} \\
& ([\mathbf{X} = \mathbf{x} \wedge Y = y_a] (Z(\mathbf{u}) = z_c) \wedge [\mathbf{X} = \mathbf{x} \wedge Y = y_b] (Z(\mathbf{u}) = z_d)) \\
& \text{(Affect-C)}^{15}
\end{aligned}$$

THEOREM 3.3.2. *(Affect) is equivalent to (Affect-C).*

¹⁴The proof of equivalence between (NC) and (GC5) is extremely similar to the proof given in (G2). I leave the proof to the reader.

¹⁵The C in (Affect-C) denotes the contrastive intervention between $Y = y_a$ and $Y = y_b$ where $a \neq b$.

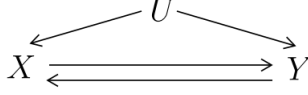


FIGURE 3.3.2

PROOF. Suppose that (Affect) is true. Let y_a be the value of Y under the intervention $[\mathbf{X}=\mathbf{x}]$, given $\mathbf{U}=\mathbf{u}$. From $[\mathbf{X}=\mathbf{x}](Z(\mathbf{u})=z_d)$ and $[\mathbf{X}=\mathbf{x}](Y(\mathbf{u})=y_b)$, $[\mathbf{X}=\mathbf{x} \wedge Y=y_b](Z(\mathbf{u})=z_d)$ follows by composition. For the sake of contradiction, suppose that $y_a=y_b$. However, given that $z_c \neq z_d$, $[\mathbf{X}=\mathbf{x} \wedge Y=y_b](Z(\mathbf{u})=z_c)$ and $[\mathbf{X}=\mathbf{x} \wedge Y=y_b](Z(\mathbf{u})=z_d)$ form a contradiction (by equality). Hence, $y_a \neq y_b$ and (Affect-C) follows from $[\mathbf{X}=\mathbf{x} \wedge Y=y_a](Z(\mathbf{u})=z_c)$ and $[\mathbf{X}=\mathbf{x} \wedge Y=y_b](Z(\mathbf{u})=z_d)$. Conversely, suppose that (Affect-C) is true. Let z_c be the value of Z under the intervention $[\mathbf{X}=\mathbf{x}]$, given $\mathbf{U}=\mathbf{u}$. Given that the value of Z under the intervention $[\mathbf{X}=\mathbf{x} \wedge Y=y_a]$ is different from the value under the intervention $[\mathbf{X}=\mathbf{x} \wedge Y=y_b]$, it follows that either $[\mathbf{X}=\mathbf{x} \wedge Y=y_a](Z(\mathbf{u}) \neq z_c)$ or $[\mathbf{X}=\mathbf{x} \wedge Y=y_b](Z(\mathbf{u}) \neq z_c)$. Therefore, (Affect) follows from (Affect-C). \square

It is notable that both generalized reversibility and recursiveness are principles with regard to no cycles of certain *dependence* relations. For generalized reversibility, the dependence relation is obviously that of counterfactual dependence, more precisely, counterfactual dependence relative to a *fixed* event; whereas for recursiveness, the dependence relation is that of the notation (Affect). It can be seen from (Affect-C) that the affect relation is defined in terms of counterfactual dependence as well, but the dependence is relative to *some* event, or some *contingencies*, instead of a fixed event in generalized reversibility. The following causal model illustrates the difference at stake more explicitly.

EXAMPLE 3.3.3. $T = \langle \mathbf{U}, \mathbf{V}, R, F \rangle$

$$\mathbf{U} = \{U\}, \mathbf{V} = \{X, Y\}$$

$$R(U) = R(X) = R(Y) = \{0, 1\}$$

$$X = U \wedge Y$$

$$Y = U \vee X$$

See $G(T)$ in Figure 3.3.2.

The causal model in this example is in $T_{unig}(S)$, that is, every submodel of T has a unique solution. Interestingly, (GC5) is true in this model, but not (C6). (C6) is violated obviously due to the cycles, for example, resulted from $X \rightsquigarrow Y$ and $Y \rightsquigarrow X$. Nevertheless, when $U = 1$, the value of Y is always 1 (unless intervened as 0), no matter what value X takes. The truth of $X \rightsquigarrow Y$ is granted when $U = 0$, that is, from $[X = 0](Y(0) = 0) \wedge [X = 1](Y(0) = 1)$.

On the contrary, when $U=0$, the value of X is always 0 (unless intervened as 1), no matter what value Y takes. It is when $U=1$ that intervening the value of Y alters the value of X . From $[Y=0](X(1)=0) \wedge [Y=1](X(1)=1)$, $Y \rightsquigarrow X$ follows.

This example immediately confirms the mentioned difference of the dependence relation between (GC5) and (C6). The circumstances under which X affects Y and that under which Y affects X are different. In the former, it is $U=0$ which make $X \rightsquigarrow Y$ true in T , whereas $U=0$ does not make $Y \rightsquigarrow X$ true in T . This is the suggested contingency which is varied but not fixed. In contrast, (GC5) merely forbid the cycle of counterfactual dependence relative to a fixed event. Putting it differently, there is no u of U and no submodel of T that there is a cyclic counterfactual dependence held between X and Y when they are taken to be families of events. Thus, the *acyclicity* of the dependence relation between (GC5) and (C6) are different.¹⁶

Manifestly, (GC5) is entailed by (C6), but not vice versa.¹⁷ Therefore, in order to avoid the peculiarity by disallowing any length of cyclic counterfactual dependence, adopting (C6) is unnecessary in spite of its sufficiency. It is more preferable to adopt (GC5).

Consequently, it is intuitive that a class of models which is sound and complete for an axiomatic system consists of (C0)-(C4), (GC5), and MP is able to rule out all cyclic counterfactual dependence.¹⁸ However, it is an open question whether there is an interesting and independent characterization of the class of models that validates (GC5). Another interesting question is what logics in the Lewis-Stalnaker framework rule out all cycles of counterfactual

¹⁶In Zhang et al. (2012), we suggest a logic which rules out all cycles of counterfactual dependence and it is weaker than $\mathbf{AX}_{rec}(S)$. This logic contains (GC5), but not (C6), and it characterizes causal models that are, so to speak, acyclic (or recursive) at a *token level*, but not necessarily at a *type level*. By contrast, $\mathbf{AX}_{rec}(S)$ characterizes causal models that are recursive at a type level. The idea of distinguishing between different levels of acyclicity (or recursiveness) is borrowed from the discussion of token causal claims and type causal claims. Indeed, whether they are reducible to each other and whether one is more fundamental are controversial debates in the literature on causation. See Hausman (2005) and Ehring (2009) for an in-depth survey of these issues.

¹⁷First, by reductio, suppose that (C6) is true and (GC5) is false. That is $[\mathbf{X}=\mathbf{x}](Y_2(\mathbf{u})=y_2)$ is false and it is true that $[\mathbf{X}=\mathbf{x}](Y_2(\mathbf{u})=y'_2)$ where $y_2 \neq y'_2 \in Y_2$. Then it follows from $[\mathbf{X}=\mathbf{x} \wedge Y_1=y_1](Y_2(\mathbf{u})=y_2)$ that Y_1 affects Y_2 . Next, either Y_2 affects Y_3 or Y_2 does not affect Y_3 . Suppose it does not, then it follows that $[\mathbf{X}=\mathbf{x}](Y_3(\mathbf{u})=y_3)$. Also, $[\mathbf{X}=\mathbf{x}](Y_4(\mathbf{u})=y_4)$ follows from $[\mathbf{X}=\mathbf{x} \wedge Y_3=y_3](Y_4(\mathbf{u})=y_4)$ and finally $[\mathbf{X}=\mathbf{x}](Y_1(\mathbf{u})=y_1)$ from $[\mathbf{X}=\mathbf{x} \wedge Y_k=y_k](Y_1(\mathbf{u})=y_1)$. But then $[\mathbf{X}=\mathbf{x} \wedge Y_1=y_1](Y_2(\mathbf{u})=y_2)$ and $[\mathbf{X}=\mathbf{x}](Y_1(\mathbf{u})=y_1)$ entail that $[\mathbf{X}=\mathbf{x}](Y_2(\mathbf{u})=y_2)$ which contradicts $[\mathbf{X}=\mathbf{x}](Y_2(\mathbf{u})=y'_2)$. Thus, Y_2 affects Y_3 . A similar argument works in showing that Y_3 affects Y_4 and so on till Y_k affects Y_1 . However, it then contradicts (C6) and hence (GC5) follows from (C6). On the other hand, a simple model like Example 2.2.2 shows that (GC5) does not entail (C6).

¹⁸It is an axiomatic system stronger than $\mathbf{AX}_{uniq}(S)$. I do not include (C5) as it follows from (GC5).

dependence. The remaining of this chapter will show that the logic **VCSR** proposed in the previous chapter is precisely one such logic.

The general argument is this: as **VCSR** contains the translated reversibility (i.e. (C5t3)) and reversibility is shown to be equivalent to the irreversibility principle (i.e. no mutual counterfactual dependence), that logic obviously rules out mutual counterfactual dependence between distinct single events. If, moreover, we can show that cyclic counterfactual dependence entails mutual counterfactual dependence in that logic, then no cyclic counterfactual dependence is allowed in **VCSR**. The core step of this argument is thus to show that given a cycle of counterfactual dependence among some distinct single events, there is a mutual counterfactual dependence between any two distinct single events in the cycle.¹⁹

Let $(E_{X_i} \rightrightarrows E_{X_j} | E_Y)$ represent the formula $((O(E_Y) \wedge O(E_{X_i})) \Box \rightarrow O(E_{X_j})) \wedge ((O(E_Y) \wedge \sim O(E_{X_i})) \Box \rightarrow \sim O(E_{X_j}))$, where $O(E_Y), O(E_{X_i}), O(E_{X_j}) \in L$. Similarly, let $(E_{X_i} \rightrightarrows E_{X_j})$ abbreviate the formula that $(O(E_{X_i}) \Box \rightarrow O(E_{X_j})) \wedge (\sim O(E_{X_i}) \Box \rightarrow \sim O(E_{X_j}))$. Naturally, $(E_{X_i} \rightrightarrows E_{X_j} | E_Y)$ is read as E_{X_j} counterfactually depends on E_{X_i} relative to E_Y . Then, (NC) becomes:

$$\begin{aligned} & \bigwedge_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} D(E_{X_i}, E_{X_j} | E_Y) \\ & \supset (\bigwedge_{1 \leq i \leq k-1} (E_{X_i} \rightrightarrows E_{X_{i+1}} | E_Y) \supset \sim (E_{X_k} \rightrightarrows E_{X_1} | E_Y)) \end{aligned}$$

Similarly, (C5t3) can be paraphrased as:

$$D(E_{X_i}, E_{X_j} | E_Y) \supset ((E_{X_i} \rightrightarrows E_{X_j} | E_Y) \supset \sim (E_{X_j} \rightrightarrows E_{X_i} | E_Y))$$

by substituting ϕ as $O(E_Y)$, ψ and χ with $O(E_{X_i})$ and $O(E_{X_j})$ respectively.

LEMMA 3.3.4. *For any $M \in \mathbf{M}_S$, where $M = \langle W, I, f \rangle$ such that for any world $w \in W$, and any set of formulas $\{\phi_1, \dots, \phi_k\} \subseteq L$ (where $k \geq 2$), if it is true that in w that ϕ_i and ϕ_j are distinct relative to $\psi \in L$ for any i and j (ranged from 1 to k where $i \neq j$), and $((\psi \wedge \phi_1) \wedge \dots \wedge ((\psi \wedge \phi_{k-1}) \Box \rightarrow \phi_k) \wedge ((\psi \wedge \phi_k) \Box \rightarrow \phi_1))$ is true in w , then for any formulas $\phi_i, \phi_j \in \{\phi_1, \dots, \phi_k\}$ which are distinct relative to ψ , $((\psi \wedge \phi_i) \Box \rightarrow \phi_j) \wedge ((\psi \wedge \phi_j) \Box \rightarrow \phi_i)$ is true in w .²⁰*

PROOF. Suppose that, for any i and j (ranged from 1 to k where $i \neq j$), ϕ_i and ϕ_j are distinct relative to ψ in w , and $((\psi \wedge \phi_1) \Box \rightarrow \phi_2) \wedge \dots \wedge ((\psi \wedge \phi_{k-1}) \Box \rightarrow \phi_k) \wedge ((\psi \wedge \phi_k) \Box \rightarrow \phi_1)$ is true in w . First of all, consider the unique world in

¹⁹As mentioned in the very beginning of this chapter, I usually consider families of events which contain exactly two members to simulate single events (with occurrence and non-occurrence as its members) for the sake of simplicity.

²⁰The following proof is inspired by Halpern (2010). Note that in Lemma 3.3.4 (and Lemma 3.3.5 as well) that M is in \mathbf{M}_S , but not restricted to its subclass \mathbf{M}_{SR} . That is, the lemma can be proved by the original Stalnaker's semantics without our additional condition (SR). Moreover, a similar proof can be done by relaxing the class of models from \mathbf{M}_S to \mathbf{M} . However, I suspect that the lemma can be proved without the *limit assumption*.

the set of the closest $\psi \wedge (\phi_1 \vee \dots \vee \phi_k)$ -worlds to w , name it w_1 . The fact that the set of the closest $\psi \wedge (\phi_1 \vee \dots \vee \phi_k)$ -worlds is non-empty is guaranteed by the initial supposition that ϕ_i and ϕ_j are distinct relative to ψ in w for any i and j . Without loss of generality, suppose that w_1 is a $\psi \wedge \phi_1$ -world (no loss of generality due to the cycle among ϕ_1, \dots, ϕ_k). I claim that w_1 is the unique world in the set of the closest $\psi \wedge \phi_1$ -worlds to w (i.e. $f(\psi \wedge \phi_1, w) = \{w_1\}$). By reductio, suppose that w_1 is not in $f(\psi \wedge \phi_1, w)$. It entails that there is a closer $\psi \wedge \phi_1$ -world to w than w_1 , call it w_2 . Of course, w_2 is also a $\psi \wedge (\phi_1 \vee \dots \vee \phi_k)$ -world. Then w_2 is closer than w_1 to w and it immediately contradicts the supposition that w_1 is the unique world in the set of the closest $\psi \wedge (\phi_1 \vee \dots \vee \phi_k)$ -worlds to w . Thus, w_1 is the unique world in the set of the closest $\psi \wedge \phi_1$ -worlds to w . By the truth of $(\psi \wedge \phi_1) \Box \rightarrow \phi_2$ in w , w_1 is a ϕ_2 -world. Note that w_1 is also the unique world in the set of the closest $\psi \wedge \phi_2$ -worlds to w . It follows from a similar argument in proving that w_1 is the unique world in the set of the closest $\psi \wedge \phi_1$ -worlds to w .

Next, consider any formulas ϕ_i and ϕ_j where i and j ranged from 1 to k and $i \neq j$. From the initial supposition on distinct formulas, ϕ_i and ϕ_j are guaranteed to be distinct relative to ψ . Then, by applying the stated argument again and again, w_1 is the unique world in the closest $\psi \wedge \phi_i$ -worlds to w and also that in the closest $\psi \wedge \phi_j$ -worlds to w . This fact is derived from the truth of $((\psi \wedge \phi_{i-1}) \Box \rightarrow \phi_i)$ and $((\psi \wedge \phi_{j-1}) \Box \rightarrow \phi_j)$ in w . So, it follows that w_1 is a ϕ_i -world and a ϕ_j -world. Hence, $((\psi \wedge \phi_i) \Box \rightarrow \phi_j) \wedge ((\psi \wedge \phi_j) \Box \rightarrow \phi_i)$ is true in w . \square

LEMMA 3.3.5. *For any $M \in \mathbf{M}_S$, where $M = \langle W, I, f \rangle$ such that for any world $w \in W$, any event E_Y , and any single events E_{X_1}, \dots, E_{X_k} (where $k \geq 2$), if $\bigwedge_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} D(E_{X_i}, E_{X_j} | E_Y)$ and $(\bigwedge_{1 \leq i \leq k-1} (E_{X_i} \rightrightarrows E_{X_{i+1}} | E_Y) \wedge (E_{X_k} \rightrightarrows E_{X_1} | E_Y))$ are true in w , then it is true in w that $D(E_{X_i}, E_{X_j} | E_Y) \supset ((E_{X_i} \rightrightarrows E_{X_j} | E_Y) \wedge (E_{X_j} \rightrightarrows E_{X_i} | E_Y))$, for any $E_{X_i}, E_{X_j} \in \{E_{X_1}, \dots, E_{X_k}\}$.²¹*

PROOF. Firstly, suppose that $\bigwedge_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} D(E_{X_i}, E_{X_j} | E_Y)$ and $(\bigwedge_{1 \leq i \leq k-1} (E_{X_i} \rightrightarrows E_{X_{i+1}} | E_Y) \wedge (E_{X_k} \rightrightarrows E_{X_1} | E_Y))$ are true in w . That is, $((O(E_Y) \wedge O(E_{X_i})) \Box \rightarrow O(E_{X_{i+1}})) \wedge ((O(E_Y) \wedge \sim O(E_{X_i})) \Box \rightarrow \sim O(E_{X_{i+1}}))$ and $((O(E_Y) \wedge O(E_{X_k})) \Box \rightarrow O(E_{X_1})) \wedge ((O(E_Y) \wedge \sim O(E_{X_k})) \Box \rightarrow \sim O(E_{X_1}))$ are true in w . Then, consider any $E_{X_i}, E_{X_j} \in \{E_{X_1}, \dots, E_{X_k}\}$, $D(E_{X_i}, E_{X_j} | E_Y)$ is guaranteed to be true in w given that $\bigwedge_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} D(E_{X_i}, E_{X_j} | E_Y)$ is true in w . Thus, by Lemma 3.3.4, the following formulas are true in w :

²¹This lemma is the core premise in the abovementioned argument. It expresses the idea that, if there is a cyclic counterfactual dependence among distinct single events E_{X_1}, \dots, E_{X_k} , then there is a mutual counterfactual dependence between any single events E_{X_i}, E_{X_j} in $\{E_{X_1}, \dots, E_{X_k}\}$.

$((O(E_Y) \wedge O(E_{X_i})) \Box \rightarrow O(E_{X_j})) \wedge ((O(E_Y) \wedge \sim O(E_{X_i})) \Box \rightarrow \sim O(E_{X_j})),$
 $((O(E_Y) \wedge O(E_{X_j})) \Box \rightarrow O(E_{X_i})) \wedge ((O(E_Y) \wedge \sim O(E_{X_j})) \Box \rightarrow \sim O(E_{X_i})).$
Hence, $((E_{X_i} \rightrightarrows E_{X_j} | E_Y) \wedge (E_{X_j} \rightrightarrows E_{X_i} | E_Y))$ is true in w . □

THEOREM 3.3.6. *(NC) is valid in \mathbf{M}_{SR} .*

PROOF. By Theorem 2.4.2, (C5t3) is valid in \mathbf{M}_{SR} . That is, for any event E_Y , and any single events E_{X_i} and E_{X_j} , $D(E_{X_i}, E_{X_j} | E_Y) \supset ((E_{X_i} \rightrightarrows E_{X_j} | E_Y) \supset \sim(E_{X_j} \rightrightarrows E_{X_i} | E_Y))$ is true in every $w \in W$, for any $M = \langle W, I, f \rangle \in \mathbf{M}_{SR}$. By reductio, suppose that $\bigwedge_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} D(E_{X_i}, E_{X_j} | E_Y)$ and $(\bigwedge_{1 \leq i \leq k-1} (E_{X_i} \rightrightarrows E_{X_{i+1}} | E_Y) \wedge (E_{X_k} \rightrightarrows E_{X_1} | E_Y))$ are also true in an arbitrary world $w \in W$. Consider single events $E_{X_i}, E_{X_j} \in \{E_{X_1}, \dots, E_{X_k}\}$ and an event E_Y . $D(E_{X_i}, E_{X_j} | E_Y)$ is true in w given the truth of $\bigwedge_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} D(E_{X_i}, E_{X_j} | E_Y)$ in w' . By Lemma 3.3.5, it follows that $((E_{X_i} \rightrightarrows E_{X_j} | E_Y) \wedge (E_{X_j} \rightrightarrows E_{X_i} | E_Y))$ is true in w . Then, it derives a contradiction with (C5t3). Thus, either $\bigwedge_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} D(E_{X_i}, E_{X_j} | E_Y)$ is false or $(\bigwedge_{1 \leq i \leq k-1} (E_{X_i} \rightrightarrows E_{X_{i+1}} | E_Y) \wedge (E_{X_k} \rightrightarrows E_{X_1} | E_Y))$ is false in w . Hence, (NC) follows. □

Given the Theorem 2.4.18 in chapter 2 that **VCSR** is a complete axiomatization for L with respect to \mathbf{M}_{SR} , (NC) is a theorem in **VCSR**.

To summarize, in order to avoid the mentioned peculiarity in Pearl's logic by disallowing any length of cyclic counterfactual dependence, adopting generalized reversibility is sufficient. Surprisingly, the equivalent form of generalized reversibility in the Lewis-Stalnaker framework (i.e. (NC)) is contained in **VCSR**, which is an extension of Stalnaker's logic which incorporates translated reversibility (i.e. (C5t3)).

Here is a slightly different way to appreciate the result. Pearl's logic, as already noted, is peculiar because it allows cycles of counterfactual dependence except those of length two. In other words, mutual counterfactual dependence has a peculiar and special status in Pearl's logic. Now we see that the status of mutual counterfactual dependence in Stalnaker's logic is also special, and perhaps no less peculiar than, but in an opposite way to that in Pearl's. In Stalnaker's logic, although cycles of counterfactual dependence are allowed in general, there is no cycle that does not contain mutual dependence. That is of course why adding a principle that forbids mutual counterfactual dependence into Stalnaker's logic is sufficient for ruling out all cycles. To put another way, counterfactual dependence becomes a transitive concept within a cycle of counterfactual dependence, yet counterfactual dependence *per se* is not transitive.

CHAPTER 4

Conclusion

This thesis offers a detailed investigation on a characteristic principle named reversibility in Pearl’s causal modeling framework, in comparison with the renowned Stalnaker-Lewis counterfactual logics. Chapter 2 suggests that Stalnaker’s logic is more analogous to Pearl’s logic than Lewis’s logic, because the principle of definiteness in Pearl’s logic is, in its translated form, valid in Stalnaker’s semantics but not in Lewis’s. Next, Pearl’s principle of reversibility, which is entailed by the principles of definiteness and equality, is discussed. Then, a question of how to translate reversibility into the language of Stalnaker-Lewis framework is raised. After finding an appropriate translation of the principle of reversibility, I show that the principle is not valid in Stalnaker’s semantics. I then study the logic resulting from adding the principle into Stalnaker’s logic, and proved its soundness and completeness with respect to a subclass of Stalnakerian models.

The most important result of this thesis is presented in chapter 3. As Zhang et al. (2012) has shown, a special case of reversibility is precisely the claim that counterfactual dependence between distinct events is irreversible. Chapter 3 extends this result, and shows that even in its full generality, the principle of reversibility is essentially a statement about (some sort of) irreversibility of (some sort of) counterfactual dependence.

Zhang et al. (2012) also argues that Pearl’s logic is peculiar as it rules out reversible or mutual counterfactual dependence between distinct events, thanks to the principle of reversibility, but allows cyclic counterfactual dependence that involves more than two distinct events.

The peculiarity raises the question on what logics rule out cyclic counterfactual dependence in general. Although a stronger Pearlian logic which contains the principle of recursiveness does, it is more than necessary. I advocate a weaker logic which incorporates a more general principle of reversibility (i.e. (GC5)), which is sufficient to avoid the peculiarity. Finally, I show that the translated form of (GC5) is already contained in the logic developed in chapter 2.

I shall end this thesis by indicating some open problems.

First, it is an open question whether there is any good reason to allow cyclic but not mutual counterfactual dependence. In Zhang et al. (2012), we attempt an explanation which might justify the peculiarity. For example, one cannot

tell from the structure of the dependence whether two families of events E_1 and E_2 are identical if they have mutual counterfactual dependence. By contrast, if one is told that families of events E_1 , E_2 , and E_3 form a cycle of counterfactual dependence that does not contain any sub-cycle of length two, the structure entails the non-identity of the events under some reasonable criterion of event identity (Indeed, one may argue that Leibniz’s Law of the indiscernibility of identicals suffice to account for the non-identity). The criterion we have in mind is a *counterfactual criterion of event identity*, which is inspired by the causal criterion of event identity due to Davidson (1969).

Another question concerns the significance of (GC5). Unlike Pearl’s elegant characterization of $\mathbf{T}_{unig}(S)$, the class of models with unique solution, by reversibility (i.e. (C5)), and that of $\mathbf{T}_{rec}(S)$, the class of models with no feedback relation, by recursiveness (i.e. (C6)), it is not obvious what particular feature the class of models characterized by (GC5) consists. One possibility is that the class of models can be characterized by some sort of *token-recursiveness*. Usually, in the literature on causation, recursiveness refers to no feedback relation among variables. But there can also be recursiveness at the value-level, or what I called token-recursiveness. For example, it is intuitively clear that John’s being late for school and his insomnia the night before do not causally influence one another, even though it is not entirely implausible that being late and insomnia have no mutual causal influence in general. Thus, (GC5) may correspond to the models which forbid the causal cyclicity of localized event-tokens. However, the details need to be further explored.

Finally, as Halpern (2010) showed, there is a formula which is valid in Stalnaker’s logic but invalid in Pearl’s logic. Together with the fact that reversibility is not valid in Stalnaker’s logic, these two logics are not *comparable*. It is natural to expect that **VCSR** is the weakest common extension of them. But a rigorous statement and proof of this conjecture has to await another occasion.

Appendix

The following Lewisian model (i.e. a model in **M**) contains no pairwise mutual (relative) counterfactual dependence, but the principle of reversibility fails to hold. Thus, it is a counter-model to the claim that irreversibility principle 3 entails the principle of reversibility in Lewis's **VC**.

$$M = \langle W, I, f \rangle$$

$$W = \{w_0, w_1, \dots, w_{17}\}$$

$$E_X = \{E_{x_0}, E_{x_1}\}, E_Y = \{E_{y_1}, E_{y_2}, E_{y_3}\}, E_Z = \{E_{z_1}, E_{z_2}, E_{z_3}\}$$

$$O(E_{x_0}), O(E_{x_1}), O(E_{y_1}), O(E_{y_2}), O(E_{y_3}), O(E_{z_1}), O(E_{z_2}), O(E_{z_3}) \in L$$

$$Iw_0(O(E_{x_0})) = Iw_0(O(E_{y_1})) = Iw_0(O(E_{z_1})) = \text{T}$$

$$Iw_1(O(E_{x_1})) = Iw_1(O(E_{y_1})) = Iw_1(O(E_{z_1})) = \text{T}$$

$$Iw_2(O(E_{x_1})) = Iw_2(O(E_{y_2})) = Iw_2(O(E_{z_2})) = \text{T}$$

$$Iw_3(O(E_{x_1})) = Iw_3(O(E_{y_2})) = Iw_3(O(E_{z_3})) = \text{T}$$

$$Iw_4(O(E_{x_1})) = Iw_4(O(E_{y_3})) = Iw_4(O(E_{z_2})) = \text{T}$$

$$Iw_5(O(E_{x_0})) = Iw_5(O(E_{y_2})) = Iw_5(O(E_{z_2})) = \text{T}$$

$$Iw_6(O(E_{x_0})) = Iw_6(O(E_{y_2})) = Iw_6(O(E_{z_3})) = \text{T}$$

$$Iw_7(O(E_{x_0})) = Iw_7(O(E_{y_3})) = Iw_7(O(E_{z_2})) = \text{T}$$

$$Iw_8(O(E_{x_1})) = Iw_8(O(E_{y_1})) = Iw_8(O(E_{z_2})) = \text{T}$$

$$Iw_9(O(E_{x_1})) = Iw_9(O(E_{y_1})) = Iw_9(O(E_{z_3})) = \text{T}$$

$$Iw_{10}(O(E_{x_1})) = Iw_{10}(O(E_{y_2})) = Iw_{10}(O(E_{z_1})) = \text{T}$$

$$Iw_{11}(O(E_{x_1})) = Iw_{11}(O(E_{y_3})) = Iw_{11}(O(E_{z_1})) = \text{T}$$

$$Iw_{12}(O(E_{x_1})) = Iw_{12}(O(E_{y_3})) = Iw_{12}(O(E_{z_3})) = \text{T}$$

$$Iw_{13}(O(E_{x_0})) = Iw_{13}(O(E_{y_1})) = Iw_{13}(O(E_{z_2})) = \text{T}$$

$$Iw_{14}(O(E_{x_0})) = Iw_{14}(O(E_{y_1})) = Iw_{14}(O(E_{z_3})) = \text{T}$$

$$Iw_{15}(O(E_{x_0})) = Iw_{15}(O(E_{y_2})) = Iw_{15}(O(E_{z_1})) = \text{T}$$

$$Iw_{16}(O(E_{x_0})) = Iw_{16}(O(E_{y_3})) = Iw_{16}(O(E_{z_1})) = \text{T}$$

$$Iw_{17}(O(E_{x_0})) = Iw_{17}(O(E_{y_3})) = Iw_{17}(O(E_{z_3})) = \text{T}$$

$$f(O(E_{x_1}), w_0) = \{w_1, w_2, w_3, w_4\}, f(O(E_{x_0}), w_0) = \{w_0\},$$

$$f(O(E_{x_1}) \wedge O(E_{y_1}), w_0) = f(O(E_{x_1}) \wedge O(E_{z_1}), w_0) = \{w_1\},$$

$$f(O(E_{x_1}) \wedge O(E_{y_2}), w_0) = \{w_2, w_3\},$$

$$f(O(E_{x_1}) \wedge O(E_{z_2}), w_0) = \{w_2, w_4\},$$

$$f(O(E_{x_1}) \wedge O(E_{y_3}), w_0) = \{w_4\}, f(O(E_{x_1}) \wedge O(E_{z_3}), w_0) = \{w_3\},$$

$$f(O(E_{x_0}) \wedge O(E_{y_1}), w_0) = f(O(E_{x_0}) \wedge O(E_{z_1}), w_0) = \{w_0\},$$

$$f(O(E_{x_0}) \wedge O(E_{y_2}), w_0) = \{w_5, w_6\},$$

$$f(O(E_{x_0}) \wedge O(E_{z_2}), w_0) = \{w_5, w_7\},$$

$$\begin{aligned}
f(O(E_{x_0}) \wedge O(E_{y_3}), w_0) &= \{w_7\}, f(O(E_{x_0}) \wedge O(E_{z_3}), w_0) = \{w_6\}, \\
f(O(E_{x_1}) \wedge O(E_{y_1}) \wedge O(E_{z_2}), w_0) &= \{w_8\}, \\
f(O(E_{x_1}) \wedge O(E_{y_1}) \wedge O(E_{z_3}), w_0) &= \{w_9\}, \\
f(O(E_{x_1}) \wedge O(E_{y_2}) \wedge O(E_{z_1}), w_0) &= \{w_{10}\}, \\
f(O(E_{x_1}) \wedge O(E_{y_3}) \wedge O(E_{z_1}), w_0) &= \{w_{11}\}, \\
f(O(E_{x_1}) \wedge O(E_{y_3}) \wedge O(E_{z_3}), w_0) &= \{w_{12}\}, \\
f(O(E_{x_0}) \wedge O(E_{y_1}) \wedge O(E_{z_2}), w_0) &= \{w_{13}\}, \\
f(O(E_{x_0}) \wedge O(E_{y_1}) \wedge O(E_{z_3}), w_0) &= \{w_{14}\}, \\
f(O(E_{x_0}) \wedge O(E_{y_2}) \wedge O(E_{z_1}), w_0) &= \{w_{15}\}, \\
f(O(E_{x_0}) \wedge O(E_{y_3}) \wedge O(E_{z_1}), w_0) &= \{w_{16}\}, \\
f(O(E_{x_0}) \wedge O(E_{y_3}) \wedge O(E_{z_3}), w_0) &= \{w_{17}\}.
\end{aligned}$$

In this model, for example, $((O(E_{x_1}) \wedge O(E_{y_1})) \Box \rightarrow O(E_{z_1}))$ and $((O(E_{x_1}) \wedge O(E_{z_1})) \Box \rightarrow O(E_{y_1}))$ are true in w_0 , yet $(O(E_{x_1}) \Box \rightarrow O(E_{y_1}))$ is false in w_0 . Note that the existence of $\{w_8, \dots, w_{17}\}$ aims to capture the antecedent of the irreversibility principle 3 that $D(E_Y, E_Z | E_{x_0})$ and $D(E_Y, E_Z | E_{x_1})$ are true in w_0 . That is, E_Y and E_Z are distinct families of events relative to E_{x_0} and also E_{x_1} .

Bibliography

- [1] Arlo-Costa, H. (2009). The logic of conditionals. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2009 ed.). Retrieved from <http://plato.stanford.edu/archives/spr2009/entries/logic-conditionals/>
- [2] Armstrong, D. (1978). *A theory of universals*. Cambridge: Cambridge University Press.
- [3] Barker, S. (1999). Counterfactuals, probabilistic counterfactuals and causation. *Mind*, 108(431), 427-469.
- [4] Bennett, J. (1984). Counterfactuals and temporal direction. *Philosophical Review*, 93(1), 57-91.
- [5] Bennett, J. (2003) . *A philosophical guide to conditionals*. Oxford: Clarendon Press.
- [6] Bigaj, T. (2004). Counterfactuals and spatiotemporal events. *Synthese*, 142(1), 1-19.
- [7] Chisholm, R. (1946). The contrary-to-fact conditional. *Mind*, 55, 289-307.
- [8] Choi, S. (2007). Causation and counterfactual dependence. *Erkenntnis*, 67, 1-16.
- [9] Davidson, D. (1969). The individuation of events. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (pp. 216-234). Dordrecht: Reidel.
- [10] Downing, P. B. (1958). Subjunctive conditionals, time order, and causation. *Proceedings of the Aristotelian Society*, 59, 125-140.
- [11] Ehring, D. (2009). Causal relata. In P. Menzies, C. Hitchcock & H. Beebe (Eds.), *The Oxford Handbook of Causation* (pp. 387-413). Oxford: Oxford University Press.
- [12] Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science*, 68(3), S313-S324.
- [13] Ellis, B., Jackson, F., & Pargetter, R. (1977). An objection to possible-world semantics for counterfactual logics. *Journal of Philosophical Logic*, 6, 355-357.
- [14] Elster, J. (1978). *Logic and society: Contradictions and possible worlds*. New York: Wiley.
- [15] Fearon, J. (1991). Counterfactuals and hypothesis testing in political science. *World Politics*, 43, 169-195.
- [16] Field, H. (2003). Causation in a physical world. In M. J. Loux & D. Zimmerman (Eds.), *The Oxford Handbook of Metaphysics* (pp. 435-460). Oxford: Oxford University Press.
- [17] Fisher, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica*, 38, 73-92.
- [18] Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3, 151-182.
- [19] Goodman, N. (1955). *Fact, fiction and forecast*. Cambridge, MA: Harvard.

- [20] Hall, N. (2000). Causation and the price of transitivity. *The Journal of Philosophy*, 97(4), 198-222.
- [21] Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12, 317-337.
- [22] Halpern, J. Y. (2010). From causal models to counterfactual structures. *Proceedings of the Twelfth International Conference on Principles of Knowledge Representation and Reasoning*, 153-60.
- [23] Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science*, 56(4), 843-887.
- [24] Handfield, T., Twardy, C. R., Korb, K. B., & Oppy, G. (2007). The metaphysics of causal models: Where's the biff? *Erkenntnis*, 68(2), 149-168.
- [25] Harel, D. (1979). *First-order dynamic logic*. Berlin & New York: Springer-Verlag.
- [26] Hausman, D. M. (1998). *Causal asymmetries*. New York: Cambridge University Press.
- [27] Hausman, D. M. (2005). Causal relata: Tokens, types, or variables? *Erkenntnis*, 63(1), 33-54.
- [28] Hendrickson, N. (2010). Counterfactual reasoning and the problem of selecting antecedent scenarios. *Synthese*. DOI 10.1007/s11229-010-9824-1.
- [29] Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, 39, 632-657.
- [30] Hiddleston, E. (2005). Causal powers. *British Journal for the Philosophy of Science*, 56(1), 27-59.
- [31] Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98(6), 273-299.
- [32] Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review*, 116(4), 495-532.
- [33] Hitchcock, C. (2009). Structural equations and causation: Six counterexamples. *Philosophical Studies*, 144(3), 391-401.
- [34] Hughes, G. E., & Cresswell, M. J. (1996). *A new introduction to modal logic*. London: Routledge.
- [35] Hume, D. (1775/1748). *An enquiry concerning human understanding*. Oxford: Clarendon Press.
- [36] Jackson, F. (1977). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, 55(1), 3-21.
- [37] Kahneman, D., & Miller, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136-153.
- [38] Lewis, D. (1973a). Causation. *Journal of Philosophy*, 70(17), 556-567.
- [39] Lewis, D. (1973b). *Counterfactuals*. Oxford: Blackwell.
- [40] Lewis, D. (1973c). Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2, 418-446.

- [41] Lewis, D. (1977). Possible-world semantics for counterfactual logics: A rejoinder. *Journal of Philosophical Logic*, 6, 359-363.
- [42] Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13, 455-476.
- [43] Lewis, D. (1986a). Postscripts to 'causation'. In D. Lewis (Ed.), *Philosophical Papers II* (pp. 171-212). Oxford: Oxford University Press.
- [44] Lewis, D. (1986b). *The plurality of worlds*. Oxford: Blackwell.
- [45] Loewer, B. (1976). Counterfactuals with disjunctive antecedents. *Journal of Philosophy*, 73(16), 531-537.
- [46] Lowe, E. J. (1995). The truth about counterfactuals. *The Philosophical Quarterly*, 45(178), 41-59.
- [47] Maudlin, T. (2007). *The metaphysics within physics*. Oxford: Oxford University Press.
- [48] Nute, D. (1975). Counterfactuals. *Notre Dame Journal of Formal Logic*, 16, 476-482.
- [49] Nute, D. (1976). David Lewis and the analysis of counterfactuals. *Noûs*, 10(3), 355-361.
- [50] Paul, L. A. (2009). Counterfactual theories. In P. Menzies, C. Hitchcock & H. Beebe (Eds.), *The Oxford Handbook of Causation* (pp. 158-184). Oxford: Oxford University Press.
- [51] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82, 669-710.
- [52] Pearl, J. (2009). *Causality: Models, reasoning, and inference*. (2nd ed.). Cambridge, UK: Cambridge University Press.
- [53] Quine, W. V. O. (1950). *Methods of logic*. New York: Holt.
- [54] Ramachandran, M. (1997). A counterfactual analysis of causation. *Mind*, 106(422), 263-277.
- [55] Ramsey, F. P. (1990/1929). General propositions and causality. In D. H. Mellor, *Philosophical Papers* (pp. 145-163). Cambridge: Cambridge University Press.
- [56] Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, 34(2), 175-221.
- [57] Schaffer, J. (2004). Counterfactuals, causal independence and conceptual circularity. *Analysis*, 64(4), 299-308.
- [58] Shoham, Y. (1990). Nonmonotonic reasoning and causation. *Cognitive Science*, 14, 213-252.
- [59] Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. (2nd ed.). Cambridge, MA: MIT Press.
- [60] Spirtes, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science*, 71(5), 833-845.
- [61] Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in Logical Theory* (pp. 98-112). Oxford: Blackwell.
- [62] Stalnaker, R., & Thomason, R. H. (1970). A semantic analysis of conditional logic. *Theoria*, 36, 23-42.
- [63] Stalnaker, R. (1975). Indicative conditionals. *Philosophia*, 5(3), 269-286.

- [64] Stalnaker, R. (1981). A defense of conditional excluded middle. In W. Harper, R. Stalnaker & G. Pearce (Eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time* (pp. 41-56). Dordrecht: Reidel.
- [65] Strotz, R. H., & Wold, H. O. A. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, 28, 417-427.
- [66] Weber, M. (1949). Objective possibility and adequate causation in historical explanation. In E. A. Shils & H. A. Finch (Eds.), *The Methodology of the Social Sciences* (pp. 50-112). New York, The Free Press.
- [67] Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford & New York: Oxford University Press.
- [68] Woodward, J. (2004). Counterfactuals and causal explanation. *International Studies in the Philosophy of Science*, 18(1), 41-72.
- [69] Zhang, J. (2011). A Lewisian logic of causal counterfactuals. *Minds and Machines*. DOI 10.1007/s11023-011-9261-z.
- [70] Zhang, J., Lam. W. Y., & de Clercq, R., (2012). A peculiarity in Pearl's logic of interventionist counterfactuals. *Journal of Philosophical Logic*, forthcoming.