Utrecht University

Artificial Intelligence

Master's Thesis

Developing Stit Models For Causal Models

Author
Nina Hendriks
5986591

Supervisors
Dr. Johannes Korbmacher
Dr. Sander Beckers

September 2, 2019

Abstract

As artificial intelligence is becoming more influential, it has become desirable to incorporate formal accounts of responsibility in techniques relying on AI. Finding a relation between stit logic and causal models would therefore be a great development, since both systems can be used for defining different parts of responsibility. Causal models are a great tool for modelling the causation part of responsibility while stit logic can be used for modelling parts of responsibility that causal models cannot effectively represent. Few people have, however, studied the relation between causal models and stit logic. The primary goal of this project will be to see whether it is possible to interpret one formalisation of responsibility in terms of the other.

Causal models use conditional probability distributions and directed graphs to model causality among variables. They are widely used in many disciplines, including artificial intelligence and philosophy. Since causality is crucial for the formalisation of responsibility, causal models can be used for this purpose. Little work has, however, been done on using causal models to formalise responsibility.

Stit logic is a logic containing the "stit" operator. "Stit" is an acronym for "see to it that" and the corresponding operator is used to model the effect that an agent has on a specified variable in the future. This project intends to find out how stit logic is related to causation by interpreting stit logic in terms of causal models.

Contents

1 Introduction **3**
- 1.1 Stit logics . . . 4
- 1.2 Causal models . . . 4
- 1.3 Goals . . . 5
- 1.4 Outline . . . 6

2 Causal Models **7**
- 2.1 Definitions . . . 7
- 2.2 Interventions . . . 8
- 2.3 Parent Sets and Timing . . . 9
- 2.4 Semantics . . . 10

3 Stit Logics **11**
- 3.1 Structure . . . 11
- 3.2 Model . . . 12
- 3.3 Semantics . . . 13
  - 3.3.1 Chellas Stit . . . 13
  - 3.3.2 Deliberative Stit . . . 13

4 Transformation λ **15**
- 4.1 $Moments_{\mathcal{M}}$ . . . 15
- 4.2 Relation $\leq_{\mathcal{M}}$ . . . 15
- 4.3 Tree $T_{\mathcal{M}}$ . . . 16
- 4.4 Histories . . . 17
- 4.5 Order and Instants . . . 17
- 4.6 Choice . . . 17
- 4.7 Stit model $\lambda(\mathcal{M})$ . . . 18
- 4.8 Properties of models in the image of λ . . . 19
- 4.9 Example . . . 20
  - 4.9.1 Transformation from $\mathcal{M}^F$ to $\lambda(\mathcal{M}^F)$ . . . 20
  - 4.9.2 Semantics . . . 21

5 Translation of Axioms **23**
- 5.1 Characterization of Interventions . . . 23
- 5.2 Language $\mathcal{L}_G^+$ . . . 24
- 5.3 Stit and Interventions . . . 25
  - 5.3.1 Chellas Stit . . . 25
  - 5.3.2 Deliberative Stit . . . 26

5.4 Axioms 26

6 Causation and Applications **28**
6.1 Counterfactual Causation 28
6.2 Halpern and Pearl 29
6.3 Deliberative stit in the image of λ 30
6.4 Rock Throwing Example 30
6.4.1 Preemtion 32
6.4.2 Counterfactual Reasoning 32
6.4.3 The Halpern and Pearl Definition 32
6.5 Application of λ 33
6.5.1 Model $\lambda(\mathcal{M})_1$ 33
6.5.2 Model $\lambda(\mathcal{M})_2$ 34
6.6 Conclusion 35

7 Conclusion **36**
7.1 Observations 36
7.1.1 Properties of models 36
7.1.2 Operators 37
7.1.3 Responsibility 37
7.2 Future research 37

Chapter 1

Introduction

Due to the rapid development of artificial intelligence, the influence of decisions made by artificial intelligence-based systems is rapidly growing. There are many situations in which technology using artificial intelligence directly affects the people it interacts with. Some examples are self-driving cars sharing the roads with human drivers, trade bots acting on the stock market, and medical diagnosis software (Broersen 2014; Matthias 2004). Since the actions performed by these systems affect people, it is desirable to have an account of responsibility that can evaluated who is responsible for actions performed by artificial intelligence-based machines.

Technology has played a role in our lives for a long time. Consequently, it is not new that there are situations in which responsibility has to be assigned for an event that was caused by a machine. As described in (Matthias 2004), often the owner, operator, or developer of the machine is held accountable for the effects of actions by performed by the machine. The development of artificial intelligence does, however, present some challenges that make this way of assigning responsibility inadequate.

The development of artificial intelligence has led to the introduction of autonomous agents. These are software entities that can make autonomous decisions (Segerberg, Meyer, and Kracht 2016). Some of these agents can learn from their environment, which means that their behaviour is not always in the control of the developer or owner of the agent. The agent's behaviour may change in ways that are unpredictable to its developers as a result of learning. Responsibility cannot be assigned the way we are used to, since no person can predict the actions of such an agent (Matthias 2004).

Consequently, due to the fast development of artificial intelligence, a more elaborate definition of responsibility is desirable. The goal of this thesis is to define such a definition in a formal language.

Several attempts have already been made at developing a formal definition of responsibility (Broersen 2014; Halpern and Kleiman-Weiner 2018; Cholvy, Cuppens, and Saurel 1997). Some of these makes use of causal models (Halpern and Kleiman-Weiner 2018). In (Broersen 2014), it is mentioned that stit models are also used for the discussion of responsibility. Both these types of models have very different properties which makes them suitable for a formalisation of responsibility for different reasons. This thesis will discuss how combining these models can further the development of a formal account of responsibility. This

chapter will discuss both types of models. It will also introduce the objective of the thesis and describe why the different models can be used to achieve this objective.

1.1 Stit logics

The first types of logic that will be discussed in this thesis is a type of logics of action. The logic of action is a group of logics that reason over actions in formal languages (Segerberg, Meyer, and Kracht 2016). The logic of action was first studied in philosophy, but has been studied in many other fields, including artificial intelligence. The logic of action is a suitable starting point for a study of responsibility in the context of artificial intelligence, since a good definition of responsibility must be able to make formal claims about actions (Cholvy, Cuppens, and Saurel 1997). A thorough description of logic of action can be found in (Segerberg, Meyer, and Kracht 2016).
The type of logic of action that we will consider for the development of a theory of responsibility is called "stit logics". Stit logics get their name from the acronym for "seeing to it that". They provides a way to study actions by means of formal languages (Segerberg, Meyer, and Kracht 2016).
This logic reasons over models called stit models. These are models with a tree-like structure of which each branch can be considered a possible flow of time. Agents can perform actions that determine trough which branch time will flow. The actual power that the agents have is limited, as is described in (Segerberg, Meyer, and Kracht 2016). Agents cannot determine precisely what flow of time will take place. They can, however, rule out certain branches every time they perform an action. This logic contains operators that can express the effects of these actions. Besides that, the logics contains temporal operators that are similar to those used in "regular" temporal logics.
Stit models have the desirable property of having a language that can express whether an agent has guaranteed some event to happen by performing an action. This way of expressing whether some outcome was guaranteed by an action of an agent can be used to built arguments for the accountability of agents.
An elaborate discussion on stit models can be found in (Belnap, Perloff, and Xu 2001) and (Segerberg, Meyer, and Kracht 2016).

1.2 Causal models

In (Halpern and Kleiman-Weiner 2018), it is stated that responsibility is a concept that should be defined based on causality, blameworthiness and intention. In stit logics, there is no obvious way to express causality. Moreover, there are no restrictions on whether the valuation of variables in stit models must respect causal laws. This is one of the reasons that a definition of responsibility in terms of stit models might feel incomplete to the authors of (Halpern and Kleiman-Weiner 2018). In this thesis, the discussion of stit models will therefore be complemented with a discussion of a different type of model that can be used to express causality. The models that will be used for this purpose are called causal models.
Causal models describe the causal relations between a set of variables. They are

described by a set of variables and a set of rules determining the causal relations between these variables. Usually, these rules are structural equations expressing the values of variables in terms of the values of other variables. Causal models have the desirable property that they provide a way to study how the values of variables affect each other. Causal models are discussed more elaborately in (Pearl 2000) and (Hitchcock 2018)
One of the most useful features of causal models are interventions. These are operations on causal models that show what the effects on the values of variables is when one variable is forced to take on some value. Interventions can be used to reason over hypothetical situations in the language of causal models. This is why causal models are suitable for reasoning over counterfactual dependency (Hitchcock 2018).
Several definitions of actual causation have been proposed that are expressed in terms of the language of causal models(Halpern and Pearl 2005; Beckers and Vennekens 2018). Many of these definition are based on the idea of counterfactual dependency. As a result, causal models played a big role in the development of definitions of actual causation.
The ability to express and reason over actual causation is an advantage of causal models over stit models. Stit theory is a logic that studies actions, and the valuations of variables in its models does not have to respect any causal laws. Since causality has been called an important component of responsibility in (Halpern and Kleiman-Weiner 2018), the causal language can be considered a valuable addition to stit models for the formulation of responsibility.

1.3 Goals

The objective of this thesis is to contribute to finding a satisfactory formal definition of responsibility. This will be done by describing a space of stit models in which each model conveys causal information from a given causal models. Subsequently, reasoning over actions in these models will be compared to existing definitions of causality and responsibility.
In this thesis, a transformation will be developed that can be applied to any causal model and returns stit models with certain desirable properties. This transformation will be designed such that the causal information conveyed in the causal model is preserved in the stit model. If the transformation is indeed designed such that it preserves causal information, studying the stit models in its image would give insight in how reasoning over causality in stit models compares to reasoning over causality in causal models. Moreover, building such a transformation will create models in which the causal effects of actions can be expressed in the stit language.
Ultimately, the goal of this project is to improve the existing definitions of responsibility. This would mean that in the self-driving car case, for example, it would no longer be unclear if the passenger of such a car is responsible for any harm the car may cause. Moreover, a definition would be expressed solely in formal language, this would mean that an autonomous agent could reason over the definition itself.

1.4 Outline

The first chapters of this thesis will give an introduction to the models that are discussed with respect to responsibility. Chapter 2 will introduce causal models and Chapter 3 will introduce stit models. In Chapter 4 a transformation will be introduced that transforms and combination of a causal model and additional information into a stit model. This transformation will be denoted by λ. Chapter 4 will conclude with a description of the characteristics of models in the image λ.
Chapter 5 will discuss what the translation of expression in the language of causal models under transformations λ looks like. An axiomatization of the class of recursive causal models will be introduced in this chapter. It will be shown what these axioms say about models in the image of λ. Finally, in Chapter 6, an example scenario will be considered. A few definitions of causation expressed in terms of the language of causal models will be used to determine which events caused each other in the example. Subsequently, transformation λ will be applied to a causal model representing the example. The semantics of the stit operators will be evaluated over the resulting stit model.

Chapter 2

Causal Models

Causality is an important concept that is related to responsibility. In order to study causality many formal definitions of causality have been developed. A lot of these definitions make use of the language of causal models (Beckers and Vennekens 2018). Moreover, the language of causal models can be used to express an axiomatization of causal reasoning (Halpern 2000). In order to properly discuss causation in the remainder of this thesis, the notion of causal models and there language must therefore first be introduced. This will be the aim of this chapter.

2.1 Definitions

Each causal model is defined by a set of variables and a set of structural equations. Causal models can be modified by operations called interventions. These concepts will be introduced in this chapter. The notation and definitions used in this chapter are based on the papers (Halpern 2000), (Halpern and Pearl 2005) and (Beckers and Vennekens 2018).
The set of variables of a causal model contains two types of elements. Each variable is either exogenous or endogenous. The set of exogenous variables of a causal model will be denoted by $\mathcal{U}$. This set contains all the variables whose value does not depend on the values of other variables in the model, but instead depend exclusively on factors outside of the model. A configuration of this set of variables is called a *context* and is represented by a vector $\vec{u}$. The endogenous variables are variables whose value is affected by the values of other variables. The set of endogenous variables a causal model is denoted by $\mathcal{V}$.
$\mathcal{R}$ is a function that is defined over the set of all the variables of a causal model. For any variable in $\mathcal{U} \cup \mathcal{V}$, $\mathcal{R}$ returns a non empty set of possible values of that variable. In this thesis it will be assumed that $\mathcal{R}$ always returns a finite set. In this thesis it will be assumed that any variable can take on at least two different values.
For a vector $\vec{X}$ of variables in $\mathcal{U} \cup \mathcal{V}$, $\mathcal{R}(\vec{X})$ denotes the cross product of the sets of possibles values of variables in $\vec{X}$.

$$\mathcal{R}(\vec{X}) = \bigtimes_{X \in \vec{X}} \mathcal{R}(X)$$

A tuple containing a set of endogenous variables, a set of exogenous variables, and a function $\mathcal{R}$ is called a signature.

Definition 2.1.1. A **signature** $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$ where U represents a set of exogenous variables, $\mathcal{V}$ represents a set of endogenous variables, and $\mathcal{R}$ is the function returning all possible values of each variable.

Besides a signature, the definition of a causal model also depends on a set of structural equations, denoted by $\mathcal{F}$. For each variable $X \in \mathcal{V}$, $\mathcal{F}$ contains an equation F_X. Equation F_X expresses the value of X in terms of the values of the variables in $\{\mathcal{V} \cup \mathcal{U}\} \backslash X$.
A signature combined with a set of structural equations over that signature defines a causal model.

Definition 2.1.2. A **causal model** $\mathcal{M}$ is described by a tuple $(\mathcal{S}, \mathcal{F})$ of a signature $\mathcal{S}$ and a set of structural equations $\mathcal{F}$.

2.2 Interventions

An intervention is an operation on a causal model that modifies its set of structural equations. An intervention is denoted by an expression of the form $\vec{X} \leftarrow \vec{x}$, where $\vec{X}$ is a vector of endogenous variables and $\vec{x}$ a vector in $\mathcal{R}(\vec{X})$. When intervention $\vec{X} \leftarrow \vec{x}$ is applied to a causal model $\mathcal{M}$, the model is transformed to $\mathcal{M}^{\vec{X} \leftarrow \vec{x}}$.
This causal model has the same variables as $\mathcal{M}$. The set of structural equations of $\mathcal{M}^{\vec{X} \leftarrow \vec{x}}$ is denoted by $\mathcal{F}^{\vec{X} \leftarrow \vec{x}}$. For every variable $Y \notin \vec{X}$, this set contains the same structural equation as $\mathcal{F}$. For each variable $X_i \in \vec{X}$, $\mathcal{F}^{\vec{X} \leftarrow \vec{x}}$ contains an equation $X_i = x_i$, where X_i is the ith variables in $\vec{X}$ and x_i is the ith variable in $\vec{x}$

Definition 2.2.1. An **intervention** is an operation on a causal model. Intervention $\vec{X} \leftarrow \vec{x}$ transforms $\mathcal{M} = (\mathcal{S}, \mathcal{F})$ into $\mathcal{M}^{\vec{X} \leftarrow \vec{x}} = (\mathcal{S}, \mathcal{F}_{\vec{X} \leftarrow \vec{x}})$.

Definition 2.2.2. Given a set of structural equations $\mathcal{F}$ and an intervention $\vec{X} \leftarrow \vec{x}$, $\mathcal{F}^{\vec{X} \leftarrow \vec{x}}$ is the set with one equation $F_Y^{\vec{X} \leftarrow \vec{x}}$ for each variable $Y \in \mathcal{V}$. These equations have the following form:

- $\forall Y \notin \vec{X}$, $F_Y^{\vec{X} \leftarrow \vec{x}} = F_Y$
- $\forall X_i \in \vec{X}$, $F_{X_i}^{\vec{X} \leftarrow \vec{x}} = x_i$

Interventions can be used to define the notions of *independence* and *recursiveness*. The following definitions of these concepts are based on the definition given in (Halpern 2000).

Definition 2.2.3. Variable X is **independent** of variable Y if given some context $\vec{u}$, the value of X remains the same when the value of Y is changed to any of its possible values.

In other words, X is independent of Y if the value of X remains the same under any intervention $\vec{Y} \leftarrow \vec{y}$.

Definition 2.2.4. A causal model $\mathcal{M}$ is **recursive** if there exists a strict total order $\prec$ of the variables in $\mathcal{V}$ such that if $X \prec Y$, then Y is independent of the value of X.

Recursive causal models can be thought of as having an acyclic structure. One important property of recursive models is that given a context, there is always a unique solution to the set of structural equations of a recursive causal model. In the remainder of this discussion all causal models are assumed to be recursive.

2.3 Parent Sets and Timing

Given a causal model $\mathcal{M}$, it is possible to define for each variable in $\mathcal{V}$ what other variables directly affect its value. A variable directly affecting the value of a variable V is called a *parent* of V. The set of all parents of a variable is called the *parent-set* of that variable. For any endogenous variable V, $pa(V)$ denotes the parent-set V.

Definition 2.3.1. Variable Y is a **parent** of variable X if there exists a configuration of $\mathcal{V}\backslash\{X, Y\}$ such that after setting all these variables to this configuration, a change in the value of Y can still cause a change in the value of X trough equation F_x.

Definition 2.3.2. For any variable $V \in \mathcal{V}$, **the parent-set** $pa(V)$ denotes the set containing all variables that are parents of V.

Note that in a total order $\prec$ of a recursive causal model, as specified in definition 2.2.4, a variable is always ranked lower than all the variables in its parent set.
Based on an order as mentioned in definition 2.2.4, variables can be assigned a value, called a timing, that carries temporal information. The definition and use of the timing here are based on the definition of timing in (Beckers and Vennekens 2018).
A timing τ assigns a natural number between 0 and some maximum value k to each endogenous variable of $\mathcal{M}$. Intuitively, for any variable $V \in \mathcal{V}$, $\tau(V)$ denotes the point in time at which V obtains its value. In order for τ to provide useful temporal information, it must meet some restrictions. If it meets these restriction, it is said to be a *valid timing*.

Definition 2.3.3. A **valid timing** of a causal model $\mathcal{M}$ is a function $\tau : \mathcal{V} \rightarrow [1, k]$ for some $k \leq |\mathcal{V}|$ that assigns a natural number between 1 and k to each variable in $\mathcal{V}$ of $\mathcal{M}$ and has the following properties;

- τ is surjective
- τ assigns values such that $\forall v_i, \forall v_j, v_i \in pa(v_j) \rightarrow \tau(v_i) \leq \tau(v_j)$

By Definition 2.2.4, a valid timing always exists for a recursive causal model. Given a recursive causal model $\mathcal{M}$, there is a total order such that $\forall X, Y \in \mathcal{V}$,$X \prec Y$ if Y is independent of X. Consider the function that assigns integers to variables in the order specified by $\prec$. This function is a valid timing.
The inverse projection function τ^{-1} is the function that takes any natural number n in the range of τ and returns the variables in $\mathcal{V}$ that have n as their

timing value. If the function τ is a valid timing, then its inverse projection is a function.

$$\tau^{-1}(x) = \{V \mid \tau(V) = x\}$$

2.4 Semantics

The semantics of causal models will be evaluated over language $\mathcal{L}_{\mathcal{M}}$ with respect to a causal model $\mathcal{M} = (\mathcal{S}, \mathcal{F})$ and context $\vec{u}$. A part of $\mathcal{L}_{\mathcal{M}}$ is the set of *primitive events* which contains formulas of the form $X = x$ where X is any variable in $\mathcal{V}$ and x is in $\mathcal{R}(X)$. $\mathcal{L}_{\mathcal{M}}$ contains elements of the form $[\vec{Y} \leftarrow \vec{y}]X = x$. In such a formula of $\mathcal{L}_{\mathcal{M}}$, $\vec{Y} \leftarrow \vec{y}$ represents an intervention and $X = x$ represents a primitive event. The vectors $\vec{Y}$ and $\vec{y}$ may be empty vectors. This is the case when non-intervened on models are considered. The complete language $\mathcal{L}_{\mathcal{M}}$ contains the following elements:

$$\phi ::= [\vec{Y} \leftarrow \vec{y}]X = x \mid \phi \wedge \phi \mid \neg\phi$$

This language contains the same elements as the language $\mathcal{L}_{uniq}$ that is described in (Halpern 2000). This language will be used later in this thesis to formalize different definitions of causality. Halpern shows that this language is as expressive as the more elaborate language $\mathcal{L}^+$ (Halpern 2000). In addition to the elements in $\mathcal{L}_{uniq}$, $\mathcal{L}^+$ contains elements of the form $[\vec{Y} \leftarrow \vec{y}]\phi$ where ϕ can be any Boolean combination of primitive events.
The semantics of $\mathcal{L}_{\mathcal{M}}$ are determined by the following rules.

$(\mathcal{M}, \vec{u}) \models X = x$	$\Leftrightarrow$	x is the value of X in the unique solution of $\mathcal{F}$ given $\vec{u}$
$(\mathcal{M}, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]X = x$	$\Leftrightarrow$	$(\mathcal{M}_{\vec{Y} \leftarrow \vec{y}}, \vec{u}) \models X = x$
$(\mathcal{M}, \vec{u}) \models \neg\phi$	$\Leftrightarrow$	$(\mathcal{M}, \vec{u}) \not\models \phi$
$(\mathcal{M}, \vec{u}) \models \phi \wedge \psi$	$\Leftrightarrow$	$(\mathcal{M}, \vec{u}) \models \phi$ and $(\mathcal{M}, \vec{u}) \models \psi$

Chapter 3

Stit Logics

Stit logics are logics that can be used to reasoning over accountability for outcomes. Stit logics contain so called *stit* operators. The name of these operators is an acronym for 'seeing to it that'. This acronym first appeared in 1957 in a paper by Kanger (Belnap, Perloff, and Xu 2001). Over the course of time, different stit operators have been proposed. All operators express a version of the idea that an agent has seen to an outcome if it has ruled out all possible worlds in which that outcome is not true.
This chapter will introduce the model used in stit logics and subsequently introduce some commonly used stit operators. Most of the definitions in this chapter will be based on the book (Belnap, Perloff, and Xu 2001).

3.1 Structure

A stit model is characterised by a structure and a valuation function. This chapter will introduce all concepts that are needed to properly define both a structure and a valuation function.
The most important component of a stit structure is its set of moments, denoted by $Moments$. Besides $Moments$, the description of a structure also includes a relation $\leq$ over $Moments$. This relation must define a tree structure over the elements in $Moments$. $\leq$ defined a tree over $Moments$ if the following two conditions are met. The first is that $\leq$ must be a partial order over $Moments$. A partial order is a relation that is reflexive, antisymmetric, and transitive. The second condition is that for any element in $Moments$, the set of elements preceding that element must be a well ordered set (Moerdijk and Oosten 2018). Let $T = \langle Moments, \leq \rangle$ be the tree defined by the elements in $Moments$ and the relation $\leq$. Given a tree T, the set $Histories$ can be defined. Each element of this set represents a set of moments that is maximal linearly ordered according to $\leq$.

Definition 3.1.1. A **history** is a maximal linearly ordered subset of $Moments$

$Histories$ contains all such possible histories. For any moment m, $H(m)$ will denote the set of all histories contain moment m. Formally, this can be described as

$$H(m) = \{h \in Histories \mid m \in h\}$$

$Instant$ is a partition of $Moments$ whose equivalence classes contain moments that are at equal depth of the tree T. Two moments are at the same depth of tree T if the same number of moments separates them from the root of T. Each equivalence class of $Instant$ is called an *instant*. The instant containing moment m is denoted by $i(m)$.
The set $Agents$ specifies which agents act on the stit model. This thesis will mostly consider single-agent models. In such models, $Agents$ has one element.
Finally, the definition of the set $Choice$ is needed to complete the definition of the structure. $Choice$ is a set containing one partition of $H(m)$ per agent in $Agents$, for each moment m. The partition of $H(m)$ for agent α is denoted by $Choice^{\alpha}_{m}$. Given such a partition, two moments are said to be $Choice^{\alpha}_{m}$-equivalent if they lie on histories that are in the same equivalence class of $Choice^{\alpha}_{m}$. The equivalence class of $Choice^{\alpha}_{m}$ that contains history h is denoted by $Choice^{\alpha}_{m}(h)$. Formally, $Choice$ is described in the following way:

$$Choice = \{Choice^{\alpha}_{m} \mid m \in Moments, \alpha \in Agents\}$$

In a single-agent model, the partition of $H(m)$ for the sole agent is simply denoted by $Choice_m$. In a single-agent model, the set $Choice$ is defined in the following way:

$$Choice = \{Choice_m \mid m \in Moments\}$$

The sets that have been defined contain enough information to fully define a stit structure.

Definition 3.1.2. A **stit structure** is a tuple $\langle Moments, \leq, Instant, Agent, Choice\rangle$ where $Moments$ and $\leq$ define a tree, $Histories$ and $Instant$ are both partitions of $Moments$ and $Choice$ is a function that maps elements in $Moment$ to partitions of $Histories$.

3.2 Model

A stit model is defined by a structure, extended with a language and a valuation function.
The language of stit models is based on a set of atomic variables. Let $\mathcal{A}$ be such a set. The valuation function of a stit model maps each atomic variables to a set *moment-history pairs*. A moment-history pair is a tuple of a moment and a history containing that moment.

Definition 3.2.1. A **stit model** is a tuple $\langle S, J\rangle$ of a stit structure and a valuation function.

The language of stit models is called $\mathcal{L}_G$ and contains the following elements:

$$\phi ::= \psi \mid \neg\phi \mid \phi \wedge \phi \mid [Was : \phi] \mid [Will : \phi] \mid [Sett : \phi]$$

$$\mid [\ d\ stit:\ \phi] \mid [\ c\ stit:\ \phi]$$

Where ψ is an atomic variable. The stit operators in this language are written without reference to an agent. This thesis will mostly consider single-agent stit models. The operators can easily be expanded to include reference to agents in case this is needed for a multi-agent model.

3.3 Semantics

The semantics of a stit model are determined in the following way. Let $\psi \in \mathcal{A}$. Then;

if $\psi \in \mathcal{A}$, then $G, m/h \models \psi$	$\Leftrightarrow$	$m/h \in J(\psi)$
$G, m/h \models \neg\phi$	$\Leftrightarrow$	$G, m/h \not\models \phi$
$G, m/h \models \phi \wedge \chi$	$\Leftrightarrow$	$G, m/h \models \phi$ and $G, m/h \models \chi$
$G, m/h \models [Was : \phi]$	$\Leftrightarrow$	there is a moment m' such that $m' < m$ and $G, m'/h \models \phi$
$G, m/h \models [Will : \phi]$	$\Leftrightarrow$	there is a moment $m' \in h$ such that $m > m'$ and $G, m'/h \models \phi$
$G, m/h \models [Sett : \phi]$	$\Leftrightarrow$	for all $h' \in H(m)$, $G, m/h' \models \phi$

The semantics of formulas of the form $\psi \in \mathcal{A}$, $\neg\phi$, and $\phi \wedge \chi$ are fairly intuitive. The Was operator acts like a regular temporal operator. $[Was : \phi]$ holds at m/h if there is some moment m' ranked lower than m such that in m'/h *phi* was true. As a result of the branching structure of $S_{\mathcal{M}}$, any moment m' ranked lower than m must be on history h. The $Will$ operator is the forward-looking equivalent of the Was operator. $[Will : \phi]$ holds at m/h if there is some moment m' on h that is ranked higher than m such that in m'/h ϕ is true. The dual of this operator will be introduced and used in Section 5.1. The *Sett* operator expresses *settledness*. A formula ϕ is settled in a moment-history pair m/h if ϕ holds at all moment-history pairs with moment m. This is sometimes referred to as *historical necessity*.
There are different kinds of stit semantics that have their own stit operators. This thesis will focus on the Chellas stit semantics and the Deliberative stit semantics. The stit operators of these semantics will now be discussed.

3.3.1 Chellas Stit

The operator $[c\ stit\ :]$, is called the Chellas operator and it has the following definition: $G, m/h \models [\ c\ stit:\ \phi]$ is true if and only if $G, m/h' \models \phi$ for each $h' \in Choice_m(h)$. This definition is based on Broersen, Herzig, and Troquard 2006. Intuitively, the Chellas semantics say that an agent has *seen to it* that some formula ϕ is true if it chooses a choice cell in m such that ϕ holds in all moment history pair with moment m and a history in the chosen choice cell. When this is the case, ϕ is guaranteed to be true after the agent chooses that cell. Therefor, it is intuitive to say that by making that choice, the agent has seen to it that ϕ holds.

3.3.2 Deliberative Stit

The $[d\ stit\ :]$ operator is called the Deliberative operator. $G, m/h \models [\ d\ stit:\ \phi]$ is true if and only if the following both hold;

- $G, m/h' \models \phi$ for each $h' \in Choice_m(h)$
- $G, m/h \not\models [Sett:\ \phi]$

These semantics are based on Belnap, Perloff, and Xu 2001.
The first of these conditions is equal to the condition for Chellas stit. Like in the

Chellas semantics, in the Deliberative stit semantics the agent has only seen to something if this was guaranteed by the choice of the agent. The second of these conditions says that an agent has only seen to ϕ at a moment-history pair if ϕ was not historically necessary at that moment-history pair. This condition can intuitively be seen as the condition that an agent has not seen to an outcome if that outcome was guaranteed to happen regardless of the choice made by the agent.
The rules for Delibrative Stit to hold are more restrictive than the rules for the Chellas Operator. Consewuenlty, if $[d\ stit :\phi$ holds in some moment-history pair m/h, so does $[c\ stit :\phi]$. The following therefore always holds:

$$[d\ stit :\ \phi] \rightarrow [c\ stit :\ \phi]$$

Moreover, as is shown by (Segerberg, Meyer, and Kracht 2016), the following also always hold:

$$[d\ stit :\ \phi] \leftrightarrow ([c\ stit :\ \phi] \wedge \neg[Sett :\ \phi])$$

$$[c\ stit :\ \phi] \leftrightarrow ([d\ stit :\ \phi] \vee [Sett :\ \phi])$$

Chapter 4

Transformation λ

This chapter will introduce a map from recursive causal models to stit models. This map, called λ, maps any combination of a recursive causal model $\mathcal{M}$, a context, and a timing of that causal model to a stit model $\lambda(\mathcal{M})$.

4.1 $Moments_{\mathcal{M}}$

Assume that a causal model $\mathcal{M}$ and a timing τ are given. Let $\mathcal{I}_\tau$ be the set that contains all possible interventions on $\mathcal{M}$ that respect the following condition:

$$\vec{X} \leftarrow \vec{x} \in \mathcal{I}_\tau \Leftrightarrow \exists k \in \mathbb{N} \text{ such that } X \in \vec{X} \Leftrightarrow \tau(X) < k$$

Each moment of the stit structure of stit model $\lambda(\mathcal{M})$ will represent a tuple consisting of an element of $\mathcal{I}_\tau$ and a vector of endogenous variables.

Definition 4.1.1. A **moment** of stit model $\lambda(\mathcal{M})$ represent a tuple $(\vec{X} \leftarrow \vec{x}, \vec{V})$ of an intervention in $\mathcal{I}_\tau$ and a vector of variables from $\mathcal{V}$.

The moment representing tuple $(\vec{X} \leftarrow \vec{x}, \vec{V})$ is denoted by $m_{[\vec{X} \leftarrow \vec{x}, \vec{V}]}$.
$Moments_{\mathcal{M}}$ is the set of moments of stit structure $S_{\mathcal{M}}$. This set contains all moments that are characterized by an intervention that is in $\mathcal{I}_\tau$ and a vector containing the variables who intuitively obtain their value after that intervention. Formally, this set can be summarized in the following way:

$$Moments_{\mathcal{M}} = \{m_{[\vec{X} \leftarrow \vec{x}, \vec{V}]} \mid \vec{X} \leftarrow \vec{x} \in \mathcal{I}_\tau \wedge \Big(v \in \vec{V} \Leftrightarrow \tau(v) = \tau_{max}(\vec{X}) + 1\Big)\}$$

4.2 Relation $\leq_{\mathcal{M}}$

The relation $\leq_{\mathcal{M}}$ is a relation over the elements of $Moments_{\mathcal{M}}$.

Definition 4.2.1. $\leq_{\mathcal{M}}$ is the relationship such that if $m_{[\vec{X} \leftarrow \vec{x}, \vec{V}]}$ and $m_{[\vec{X}' \leftarrow \vec{x}', \vec{V}']}$ are two elements of the set $Moments_{\mathcal{M}}$ of a stit model $\lambda(\mathcal{M})$, then:

$$m_{[\vec{X} \leftarrow \vec{x}, \vec{V}]} \leq_{\mathcal{M}} m_{[\vec{X}' \leftarrow \vec{x}', \vec{V}']} \Leftrightarrow \Big(\vec{X} \subseteq \vec{X}'\Big) \wedge \Big(\forall X_i \in \vec{X}, X_i' \in \vec{X}', X_i = X_i' \Rightarrow x_i = x_i'\Big)$$

Intuitively, this means that moment $m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}$ is ranked lower then moment $m_{[\vec{X}'\leftarrow\vec{x}',\vec{V}']}$ if the variables in $\vec{X}$ are in $\vec{X}'$ and the interventions of both moments agree on their value assignments to variables in $\vec{X}$. When the interventions characterizing two moments do not agree on the value assignments to the variables in $\vec{X}$, these moment are incomparable. Moreover, if two moments are incomparable, this means that their interventions do not agree on the value assignment of at least one variable.

4.3 Tree $T_{\mathcal{M}}$

A tree is a partial ordered set with the property each set of elements preceding some element must be a well ordered set (Moerdijk and Oosten 2018). We will show that according to this definition, $T_{\mathcal{M}} = \langle Moments_{\mathcal{M}}, \leq_{\mathcal{M}} \rangle$ is a tree.

Theorem 1. $T_{\mathcal{M}} = \langle Moments_{\mathcal{M}}, \leq_{\mathcal{M}} \rangle$ is a tree

Proof. It will first be shown that $\leq_{\mathcal{M}}$ is a partial order over $Moments_{\mathcal{M}}$. In order to show this, it must be shown that the relation $\leq_{\mathcal{M}}$ over $Moments_{\mathcal{M}}$ is reflexive, antisymmetric and transitive.
A relation is reflexive over a set if any element of that set is related to itself. By the definition of set-inclusion, the vector over which the intervention of any moment ranges is always contained in itself. The intervention of any moment obviously assigns the same values to each variable as itself. Therefore, any moment is related to itself and $\leq_{\mathcal{M}}$ is reflexive over $Moments_{\mathcal{M}}$.
A relation is antisymmetric over a set of moments if any two moment that precede each other are identical. Let us assume that $Moments_{\mathcal{M}}$ contains two elements m and m' and these moments precede each other. Then, the vectors over which their interventions range include in each other. In other words, they are identical. Since these moments are related, their interventions assign the same values. Consequently, these moments must be identical and $\leq_{\mathcal{M}}$ is antisymmetric over $Moments_{\mathcal{M}}$.
The relation $\leq_{\mathcal{M}}$ is transitive over $Moments_{\mathcal{M}}$ if for any three elements m, m', and m'' of $Moments_{\mathcal{M}}$, $m \leq_{\mathcal{M}} m'$ and $m' \leq_{\mathcal{M}} m''$ implies that $m \leq_{\mathcal{M}} m''$. Assume that three such moments exist and $m \leq_{\mathcal{M}} m'$ and $m' \leq_{\mathcal{M}} m''$. The vectors over which the interventions of these elements range are included in each other as determined by $\leq_{\mathcal{M}}$. By the transitivity of set inclusion, the vector of the intervention of m is included in the vector of the intervention of m''. Moreover, because both moments are related to m', they must agree on the value assignment to the variables in the vector of the intervention of m. Consequently m and m'' are related and $\leq_{\mathcal{M}}$ is transitive over $Moments_{\mathcal{M}}$.
It can now be deduced that the relation $\leq_{\mathcal{M}}$ is a partial ordering of $Moments_{\mathcal{M}}$. The second condition for $\langle Moments_{\mathcal{M}}, \leq_{\mathcal{M}} \rangle$ to be a tree is that any set of moments preceding some element of $Moments_{\mathcal{M}}$ must be a well ordered set. A set is well ordered if $\leq_{\mathcal{M}}$ is a total order over that set and every subset of that set has a least element (Moerdijk and Oosten 2018).
Let $m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}$ be an arbitrary element in $Moments_{\mathcal{M}}$ and consider the set $Moments'_{\mathcal{M}} = \{m \in Moments_{\mathcal{M}} \mid m \leq m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}\}$. It will be shown that $\leq_{\mathcal{M}}$ is a well order over $Moments'_{\mathcal{M}}$.
We will first show that $\leq_{\mathcal{M}}$ is a total order over this set. A total order is a

partial order in which all elements are related. Let m' and m'' be two moments in $Moments'_{\mathcal{M}}$. As a result of the definition of $\mathcal{I}_\tau$, the vector of the intervention corresponding to one of the moments m and m' must be included in the vector of the intervention corresponding to the other moment. Since both moments are related to m, they agree on the value assignments of smallest vector among the vectors of their interventions. Consequently these moments are related.
In order to show that $Moments'_{\mathcal{M}}$ is well ordered, it must also be true that each of its subsets has a least element. Let $Moments''_{\mathcal{M}}$ be a subset of $Moments'_{\mathcal{M}}$. As a result of the definition $\mathcal{I}_\tau$, $Moments''_{\mathcal{M}}$ has finitely many elements. $\leq_{\mathcal{M}}$ is a total order over this set and the least element of $Moments''_{\mathcal{M}}$ is simply the moment that is ranked lower than all other moments.
$Moments'_{\mathcal{M}}$ must therefore be a well ordered set. Consequently, $T_{\mathcal{M}} = \langle Moments_{\mathcal{M}}, \leq_{\mathcal{M}} \rangle$ is a tree. □

4.4 Histories

The set $Histories_{\mathcal{M}}$ contains all branches of $T_{\mathcal{M}}$. By the definition of stit trees, $H(m_{[\vec{X}\leftarrow\vec{x},\vec{V}]})$ contains all histories containing moment $m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}$. This can intuitively be thought of as the set containing all branches of T going trough $m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}$.
If $h \in Histories$ branch of T, and $m_{[\vec{X}\leftarrow\vec{x},\emptyset]}$ is its maximal element, then $\vec{X} \leftarrow \vec{x}$ is called the *characterizing intervention* of h. The intervention corresponding to any moment in h is a restriction of the characterizing intervention of h.

4.5 Order and Instants

Based on the definition of the elements in $Moments$, an integer can be assigned to each moment based on its depth in tree T. For each moment m, integer $O(m)$ will denote the *order* of moment m. The notion of order is not part of the definition of stit models in (Belnap, Perloff, and Xu 2001). Intuitively the order of a moment m represents the number of times an agent must intervene on $\mathcal{M}$ before reaching moment m.

Definition 4.5.1. For any $m_{[\vec{X}\leftarrow\vec{x},\vec{V}]} \in Moments$, the **order** of $m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}$ is $O(m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}) = \max(\{\tau(X) \mid X \in \vec{X}\})$.

As a result of this definition of orders, two moments are in the same *instant*, as defined in Section 3.1, if and only if they have the same order.

4.6 Choice

As defined in Section 3.1, a single-agent stit structure must contain a choice set that contains a partition of $H(m)$ for each element moment m in its stit structure.
$Choice_{\mathcal{M}}$ will be the set that contains a partition $Choice_{m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}}$ of $H(m_{[\vec{X}\leftarrow\vec{x},\vec{V}]})$ for each element $m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}$ of $Moments_{\mathcal{M}}$. In each partition in $Choice$, all histories whose characterizing intervention assign the same value to $\vec{V}$ are in

the same equivalence class. The equivalence class containing histories whose characterizing interventions assign values $\vec{v}$ to variables $\vec{V}$ will be denoted by $Choice^{\vec{Y}\leftarrow\vec{y}}_{m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}}$ where $\vec{Y} = \vec{X} \cup \vec{V}$ and $\vec{y} = \vec{x} \cup \vec{v}$. Any moment on a history in this class must represent an intervention that assigns the values in $\vec{y}$ to the variables in $\vec{Y}$. This is a result of the fact that no intervention can be undone in a stit model in the images of λ. In this thesis, only causal models will be considered whose variables can take on at least two values. As a result, all partitions in $Choice_{\mathcal{M}}$ of any stit model $\lambda(\mathcal{M})$ contain at least two equivalence classes.
As described in Section 3.1, the equivalence class of $Choice_{m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}}$ containing history h is denoted by $Choice_{m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}}(h)$. When the semantics of $\lambda(\mathcal{M})$ are discussed it will become clear that $h \in Choice^{\vec{Y}\leftarrow\vec{y}}_{m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}}$ if and only if $\vec{Y} = \vec{y}$ holds at $m_{[\vec{X}\leftarrow\vec{x},\vec{V}]}/h$.
It should be noted that according to this definition, $Choice_{\mathcal{M}}$ contains one partition for each moment. This is enough to define a single-agent stit model. The model can easily be extended to a multi-agent model by defining one partition per moment per agent.

4.7 Stit model $\lambda(\mathcal{M})$

In the previous sections some sets were discussed that can be defined based on the information of some causal model $\mathcal{M}$, and and a timing τ. Let $Agent_{\mathcal{M}}$ be the set containing one agent α.

Definition 4.7.1. $S_{\mathcal{M}}$ is the stit structure described by $\langle Moments_{\mathcal{M}}, \leq_{\mathcal{M}}, Instant_{\mathcal{M}}, Agent_{\mathcal{M}}, Choice_{\mathcal{M}}\rangle$

$S_{\mathcal{M}}$ is the stit structure of stit model $\lambda(\mathcal{M})$. Given a stit structure $S_{\mathcal{M}}$ and a context $\vec{u}$ of causal model $\mathcal{M}$, it is possible to define stit model $\lambda(\mathcal{M})$. This will be done by taking the stit structure $S_{\mathcal{M}}$ and adding a language, valuation function, and semantics to that structure.
The language that will be evaluated with respect to the model $\lambda(\mathcal{M})$ is $\mathcal{L}_G$, which was introduced in Section 2.4. The set of atomic variables $\mathcal{A}$ that is needed to define the elements of $\mathcal{L}_G$ will contain primitive events, as defined in Section 2.4. This set contains elements of the form $X = x$ where X is an endogenous variable of $\mathcal{M}$ and $x \in \mathcal{R}(X)$ is a possible value of X.
The valuation function of $\lambda(\mathcal{M})$ maps any primitive event to a set of *moment-history pairs*. This function maps each primitive event of the form $X = x$ to all moment-history pairs that intuitively *represent* causal models in which $X = x$ holds. A moment-history pair m/h is said to represent causal model $\mathcal{M}_{\vec{X}\leftarrow\vec{x}}$ if the interventions represented by the choices that the agents had to make to reach m/h accumulate to $\vec{x} \leftarrow \vec{x}$.

Definition 4.7.2. Valuation function $J_{\mathcal{M}}$ is a function that maps any primitive event $X = x$ to moment-history pairs in the following set:

$$\{m/h \mid h \in Choice^{\vec{Y}\leftarrow\vec{y}}_{m} \text{ and } (\mathcal{M}_{\vec{Y}\leftarrow\vec{y}}, \vec{u}) \models X = x\}$$

Definition 4.7.3. Model $\lambda(\mathcal{M})$ is defined by tuple $\langle S_{\mathcal{M}}, J_{\mathcal{M}}\rangle$

The semantics of a stit model $\lambda(\mathcal{M})$ behave according to the semantics of stit models defined in Section 3.3. Note that as a result of this definition and the definition of the valuation function J, the following always holds:

$$\forall \vec{Y} \in \mathcal{V}, \forall \vec{y} \in \mathcal{R}(\vec{Y}), (h \in Choice_m^{\vec{Y} \leftarrow \vec{y}} \Leftrightarrow \lambda(\mathcal{M}), m/h \models \vec{Y} = \vec{y})$$

.

4.8 Properties of models in the image of λ

Transformation λ is not a surjection. This means that there are some stit models for which there is no combination of a causal model, context and timing that are mapped to that stit model by λ. All the stit models that are reached by λ for some combination of a causal model, timing and context share certain properties. These properties are the result of the fact that λ preserves the causal information that is in causal models. In this section some of these properties will be discussed.
Let $\mathbb{S}$ be the space of all stit models, and $\mathbb{S}_\lambda$ the subspace of $\mathcal{S}$ that contains all stit models reached by λ. Off course, all models in $\mathbb{S}_\lambda$ share their language and valuation function, as described in Section 4.7. Moreover, models in $\mathbb{S}_\lambda$ share structural properties. An important structural feature of models in the image of λ is that for every agent, moment-history pairs that are in the same instant have a choice function of the same size. An agent intervenes on variables in the order of their fixed timing. As a result, the variables that an agent can intervene on are the same for all moment-history pairs in the same instant. The size of a choice function is based on the amount of values that can be assigned to the variables that the agent can intervene on in a moment-history pair. Consequently, moment-history pairs that are in the same instant have a choice function of the same size. As a result of this, the structure of models in $\mathbb{S}_\lambda$ looks symmetrical. A consequence of this feature is that models in $\mathbb{S}_\lambda$ only represent situations in which the amount of choices available to an agent does not depend on its previous actions.
Another distinguishing feature of models in $\mathbb{S}_\lambda$ is that in each moment-history pair of these models, some atomic variable becomes true that will remain true in the remainder of the history. This is a result of the fact that interventions cannot be changed back in these models. Every choice made by an agent in such a model, corresponds to an intervention on at least on variable. This variable takes on a value as a result of that intervention. This can be expressed in that form of a primitive event, which are the atomic variables of models in $\mathbb{S}_\lambda$.
The last feature of models in $\mathbb{S}_\lambda$ that will be discussed in the section is their valuation function. In general, there are no restrictions to what a valuation function should look like for a stit model. In stit models in $\mathbb{S}_\lambda$, primitive events are true in moment-history pairs based on the valuation of the causal model represented by that moment history pair. As a result, their value depends on the context, structural equations, and interventions performed up to the moment-history pair. This is how the causal information from the causal model is preserved in the stit model.
An example of a stit model that is not in $\mathbb{S}_\lambda$ is not hard to find. Imagine a stit model describing the options of a car navigating in traffic. In this example, when the car leaves its spot it can either go left or right. When the car goes

left, it reaches a three-way junction. When is goes right it reaches a four-way junction. Imagine a stit model in which each moment represents an intersection. The choice function for the agent in each moment has one cell for each road the agent can enter at the intersection. Let the model have one variable, representing the location of the car.
There are a few reasons why this model cannot be reached by map λ. Firstly, the amount of choice cells is not the same in each instant of this model. The size of the choice partition for each moment depends on previous choices made by the agent. Moreover, the only variable of this models changes after each moment-history pair whereas in all models in $\mathbb{S}_\lambda$ after each choice, one variable takes on a permanent value. This shows that there is no causal model $\mathcal{M}$ such that $(\mathcal{M})$ would represent the situation described above.

4.9 Example

Transformation λ will be illustrated by applying it to an example of a causal model. The following example is taken from the paper (Halpern and Pearl 2005) This example concerns a recursive causal model representing the effects of rain in April and electrical storms in May and June on the occurrence of forest fires in May and June. This model will be denoted by $\mathcal{M}^F$. The variables of $\mathcal{M}^F$ are AS, ES, and F. The variable AS represents whether or not there were showers in April. This variable can take on the values 0 and 1. ES is the variable representing whether or not there were electrical storms in the months May and June. $ES = (0,0)$ and $ES = (1,1)$ represent situations in which there were storms in neither and both months respectively. $ES = (1,0)$ represents a situation in which there were electrical storms in May only and $ES = (0,1)$ represents the situation in which there were electrical storms in June only. The variable F represent whether and when a forest fire occurs in the forest. $F = 0$ is true if no fire takes place in the forest. $F = 1$ and $F = 2$ are true if there is a fire in May or June respectively.
The parents of variable F are AS and ES. AS and ES are exogenous variables. The value of F can be computed from the values of AS and ES. When $AS = 0$, a fire takes place in the first months in which there are electrical storms. When $AS = 1$, a fire only takes place in June if there are electrical storms in June. When $AS = 1$ and there are no electrical storms in June, not fire will take place.

4.9.1 Transformation from $\mathcal{M}^F$ to $\lambda(\mathcal{M}^F)$

Transformation λ will now be illustrated by describing stit model $\lambda(\mathcal{M}^F)$. Let $\vec{u}$ be the context under which AS and F are set to 0 and ES is set to $(0,0)$. Let τ be the timing such that $\tau(AS) = 1$, $\tau(ES) = 2$, and $\tau(F) = 3$. It is easy to see that this is a valid timing.
Stit model $\lambda(\mathcal{M}^F)$ will consist of structure $S_{\mathcal{M}^F}$ and valuation function $J_{\mathcal{M}^F}$. Structure $S_{\mathcal{M}^F}$ will have the set of moments as described in Section 3.1. Some examples of elements of this set are $m_{[\vec{\emptyset} \leftarrow \vec{\emptyset}, AS]}$, $m_{[AS \leftarrow 0, ES]}$, $m_{[AS \leftarrow 1, ES]}$ and $m_{[\langle AS \leftarrow 1, ES \leftarrow (0,1) \rangle, F]}$. The ordering over these moments, $\leq_{\mathcal{M}^F}$, behaves as defined in Section 3.1. An example of a branch of the tree defined by $Moments_{\mathcal{M}^F}$ and $\leq_{\mathcal{M}}$ is the following :

$$\{m_{[\vec{\emptyset} \leftarrow \vec{\emptyset}, AS]}, m_{[AS \leftarrow 1, ES]}, m_{[\langle AS \leftarrow 1, ES \leftarrow (1,1), F]}, m_{[\langle AS \leftarrow 1, ES \leftarrow (1,1), F \leftarrow 0 \rangle, \vec{\emptyset}]}\}$$

$Histories_{\mathcal{M}^F}$ is the set containing all such branches. There are twenty-four branches in this set. The set $Choice_{\mathcal{M}^F}$ will be the set as expected from the definitions in Section 3.1. The choice partition of moment $m_{[AS\leftarrow 1,ES]}$ will be specified to illustrate what the choice partitions formed under λ look like.
In partition $Choice_{m_{[AS\leftarrow 1,ES]}}$, histories are divided based on their value assignment to ES. This variable can take on four different values, and consequently there are four different equivalence classes in $Choice_{m_{[AS\leftarrow 1,ES]}}$. There are twelve histories in $H(m_{[AS\leftarrow 1,ES]})$. Those will be referred to by their characterising intervention in the remainder of this discussion. A history with characterising intervention $\vec{X}\leftarrow\vec{x}$ will be denoted by $h^{\vec{X}\leftarrow\vec{x}}$.
$Choice_{m_{[AS\leftarrow 1,ES]}}$ contains the following equivalence classes:

$$Choice^{ES\leftarrow(0,0)}_{m_{[AS\leftarrow 1,ES]}} = \{h^{\langle AS\leftarrow 1,ES\leftarrow(0,0),F\leftarrow 0\rangle}, h^{\langle AS\leftarrow 1,ES\leftarrow(0,0),F\leftarrow 1\rangle}, h^{\langle AS\leftarrow 1,ES\leftarrow(0,0),F\leftarrow 2\rangle}\}$$

$$Choice^{ES\leftarrow(0,1)}_{m_{[AS\leftarrow 1,ES]}} = \{h^{\langle AS\leftarrow 1,ES\leftarrow(0,1),F\leftarrow 0\rangle}, h^{\langle AS\leftarrow 1,ES\leftarrow(0,1),F\leftarrow 1\rangle}, h^{\langle AS\leftarrow 1,ES\leftarrow(0,1),F\leftarrow 2\rangle}\}$$

$$Choice^{ES\leftarrow(1,0)}_{m_{[AS\leftarrow 1,ES]}} = \{h^{\langle AS\leftarrow 1,ES\leftarrow(1,0),F\leftarrow 0\rangle}, h^{\langle AS\leftarrow 1,ES\leftarrow(1,0),F\leftarrow 1\rangle}, h^{\langle AS\leftarrow 1,ES\leftarrow(1,0),F\leftarrow 2\rangle}\}$$

$$Choice^{ES\leftarrow(1,1)}_{m_{[AS\leftarrow 1,ES]}} = \{h^{\langle AS\leftarrow 1,ES\leftarrow(1,1),F\leftarrow 0\rangle}, h^{\langle AS\leftarrow 1,ES\leftarrow(1,1),F\leftarrow 1\rangle}, h^{\langle AS\leftarrow 1,ES\leftarrow(1,1),F\leftarrow 2\rangle}\}$$

A few observations can be made about this partition. All histories in $H(m_{[AS\leftarrow 1,ES]})$ have a characterizing intervention that assign the same value to AS. This is a result of the no-backwards-branching property of stit structures. Moreover, all histories that are in the same equivalence class have characterizing interventions that assign the same value to ES. This is a result of the definition of choice classes in models in the image of λ. In these models, each choice equivalence class represents an intervention.
The set $Instant_{\mathcal{M}^F}$ is defined as expected from the definition in Section 3.1. $Agents_{\mathcal{M}^F}$ is the set containing the single agent α. Let $S_{\mathcal{M}^F}$ and $J_{\mathcal{M}^F}$ be the structure and valuation function as defined earlier. $\lambda(\mathcal{M}^F)$ is the model defined by this structure and valuation function.

4.9.2 Semantics

A few examples of expression in the language $\mathcal{L}_G$ that hold in moment-history pairs of $\lambda(\mathcal{M}^F)$ will be discussed. Let $h^{**} = h^{\langle AS\leftarrow 1,ES\leftarrow(1,1),F\leftarrow 0\rangle}$. The following expression holds:

$$\lambda(\mathcal{M}^F), m_{[\langle AS\leftarrow 1,ES\leftarrow(1,1)\rangle,F]}/h^{**} \models F = 2$$

This expression corresponds to the situation discussed in (Halpern and Pearl 2005). In this example, there are showers in April and subsequently there are electrical storms in May and June. As discussed in Section 3.3, the values of the variables in $m_{[\langle AS\leftarrow 1,ES\leftarrow(1,1)\rangle,F]}/h^{**}$ depend on the values of the variables in $\mathcal{M}^F_{\langle AS\leftarrow 1,ES\leftarrow(1,1)\rangle}$ with context $\vec{u}$. This is the causal model in which AS has been set to 1 and ES has been set to $(1,1)$. As discussed by Halpern and Pearl, in this causal model variable F takes on value 2.
Let $h^* = h^{\langle AS\leftarrow 0,ES\leftarrow(1,1),F\leftarrow 1\rangle}$ and consider the following expression:

$$\lambda(\mathcal{M}^F), m_{[\vec{\emptyset}\leftarrow\vec{\emptyset},AS]}/h^* \models [Sett : F = 0]$$

As discussed in the section on the semantics of this models, expressions of the form $[Sett : \phi]$ only hold in a moment-history m/h pair if ϕ holds in all moment-history pair containing m. In the moment-history pair considered here, no intervention has taken place yet but the agent will intervene on AS. Under context $\vec{u}$, before any interventions, all the variables of $\mathcal{M}^F$ have value 0. In this setting of $\mathcal{M}^F$, intervening on AS does not affect the values of other variables. Consequently, in all moment-history pairs containing $m_{[\vec{\emptyset}\leftarrow\vec{\emptyset},AS]}$, $F = 0$ is true and as a result $[Sett : F = 0]$ holds at $m_{[\vec{\emptyset}\leftarrow\vec{\emptyset},AS]}/h^*$.

Chapter 5

Translation of Axioms

The aim of this chapter will be to introduce expressions that can be used to translate expression in the language $\mathcal{L}_{\mathcal{M}}$ to expression in the language $\mathcal{L}_G$. This will be done by introducing an expression in the language $\mathcal{L}_G^+$ that characterizes moment-history pairs that represent intervention from the language $\mathcal{L}_{\mathcal{M}}$. This expression will then be used to translate the axioms in an axiomatization of the class of recursive causal models to expression $\mathcal{L}_G^+$.

5.1 Characterization of Interventions

In this section the relation between interventions on $\mathcal{M}$ and moment-history pairs of $\lambda(\mathcal{M})$ will be discussed. A moment-history pair m/h of a stit model $\lambda(\mathcal{M})$ is said to *represent* intervention $\vec{X} \leftarrow \vec{x}$ on $\mathcal{M}$ if and only if $h \in Choice_m^{\vec{X} \leftarrow \vec{x}}$. In this section it will be shown that $[d\ stit:\ \mathbf{G}\ (\vec{X} = \vec{x})]$ in a moment-history pair m/h if and only if moment-history pair m/h represents intervention $\vec{X} \leftarrow \vec{x}$. In this expression, $\mathbf{G}$ should be interpreted as the dual of the *Will* operator that was introduced in the discussion of the semantics of stit models. $\lambda(\mathcal{M}), m/h \models \mathbf{G}\ \phi$ holds whenever for any $m' \in h$, if $m' \geqslant m$, then $\lambda(\mathcal{M}), m'/h \models \phi$.

Proof. It will now be shown that $[d\ stit:\ \mathbf{G}\ (\vec{X} = \vec{x})]$ holds at the moment-history pair representing intervention $\vec{X} \leftarrow \vec{x}$. Let m/h be a moment-history pair representing intervention $\vec{X} \leftarrow \vec{x}$ in a stit model $\lambda(\mathcal{M})$. Any history h' that is $Choice_m$-equivalent to h is in $Choice_m^{\vec{X} \leftarrow \vec{x}}$. Consequently $\vec{X} = \vec{x}$ holds in m/h' and all moment-history pairs choice equivalent to m/h. Since no intervention can be undone in a stit model in the image of λ, $\vec{X} = \vec{x}$ is true in all moment-history pairs a history $Choice_m$-equivalent to h and a moment ranked higher than m. Consequently, $\mathbf{G}\ (\vec{X} = \vec{x})$ holds in all moment-history pairs choice equivalent to m/h. As a result, $[c\ stit:\ \mathbf{G}\ (\vec{X} = \vec{x})]$ holds in m/h.
As a result of the assumption that any variable of $\mathcal{M}$ can take on at least two values, $Choice_m$ contains at least one class that is not equal to $Choice_m^{\vec{X} \leftarrow \vec{x}}$. Let $Choice_m^{\vec{X} \leftarrow \vec{x}^*}$ be such a class. Let $h^* \in Choice_m^{\vec{X} \leftarrow \vec{x}^*}$. It is clear that $\vec{X} = \vec{x}^*$ holds at m/h^* and since $\vec{x} \neq \vec{x}^*$, $\mathbf{G}\ (\vec{X} = \vec{x})$ does not holds at this moment-history pair. This shows that $\lambda(\mathcal{M}), m/h \models \neg[Sett:\ \mathbf{G}\ (\vec{X} = \vec{x})]$. It has now

been shown that both $[c \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$ and $\neg[\textit{Sett : } \mathbf{G}\ (\vec{X} = \vec{x})]$ hold at m/h. By the rules of the semantics of delibaretive stit, $[d \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$ holds at any moment-history pair representing $\vec{X} \leftarrow \vec{x}$.
It will now be shown that $[d \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$ does not hold in moment-history pairs that do not represent intervention $\vec{X} \leftarrow \vec{x}$. We will consider the cases in which m/h is a moment-history pair happening before, after, or at the same instant as the intervention on $\vec{X}$ and shown that if m/h does not represent $\vec{X} \leftarrow \vec{x}$, $[d \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$ does not hold at m/h.
Assume m/h is in an **earlier** instant than the intervention on $\vec{X}$. This means that there are some variables in $\vec{X}$ that have not been intervened on yet after the agent chooses $Choice_m(h)$. As a result of the construction of $\lambda(\mathcal{M})$, there is some $h' \in Choice_m(h)$ and some $m' \geqslant m$ such that m'/h' is a moment-history pair in which $\vec{X} \neq \vec{x}$. This is the moment-history pair in which the agent intervenes on the highest ranked variables in $\vec{X}$ and assigns values that are not in $\vec{x}$. Consequently, in m/h', $\mathbf{G}\ (\vec{X} = \vec{x})$ does not hold. Since h' is choice equivalent to h in m, $[d \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$ does not hold in m/h.
Assume m/h is in an **later** instant than the intervention on $\vec{X}$. Then the intervention that happens on $\vec{X}$ in an earlier moment on m either assigned the values in $\vec{x}$, or some other set of values. If the intervention assigned the values in $\vec{x}$, $\vec{X} = \vec{x}$ holds in all moment-history pairs containing m. This remains true in all later moment-history pairs, for interventions cannot be undone in the stit models that we consider. As a result $\neg[\textit{Sett : } \mathbf{G}\ (\vec{X} = \vec{x})]$ is not true at m/h and consequently neither is $[d \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$. If the intervention assigned the values in some vector $\vec{x}' \in \mathcal{R}(X)$ that is distinct from $\vec{x}$, this means that $\vec{X} = \vec{x}'$ holds in all moment-history pairs of m, since interventions cannot be changed back. As a result, $\vec{X} = \vec{x}$ holds in none of the moment-history pairs in m and consequently neither does $[d \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$.
Finally, assume that m/h is **in the same instant** as the moment representing $\vec{X} \leftarrow \vec{x}$. We have assumed that m/h does not represent this intervention. Consequently, there must be some vector $\vec{x}'' \neq \vec{x}$ such that $h \in Choice_m^{\vec{X} \leftarrow \vec{x}''}$. By the semantics of stit models in the image of λ, it can be derived that $\vec{X} = \vec{x}''$ holds at m/h and consequently neither does $[d \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$.
It has been shown that $[d \textit{ stit : } \mathbf{G}\ (\vec{X} = \vec{x})]$ holds exclusively in moment-history pairs representing the intervention $\vec{X} \leftarrow \vec{x}$. □

5.2 Language $\mathcal{L}_G^+$

The language that is used in this thesis to reason over causal models, $\mathcal{L}_\mathcal{M}$, is built up of interventions and primitive events. Primitive events are used as the atomic variables in the language $\mathcal{L}_G$. We can therefore use the language of stit models to reason over primitive events in the causal models represented by moments in the models returned by λ.
In the previous section it was shown that moments representing some intervention can be characterized in the stit language $\mathcal{L}_G$. This allows us to reason over expressions in terms of interventions in the language $\mathcal{L}_G$. In order to do this a new operator, the intervention operator, will now be introduced. The intervention operator has the from $[\vec{X} \leftarrow \vec{x}]_\lambda$ where $\vec{X} \leftarrow \vec{x}$ is an intervention on $\mathcal{M}$.

In order to use this operator, in the remainder of this thesis stit models in the space of λ will be evaluated using a new language $\mathcal{L}_G^+$. This language contains elements of the following form:

$$\phi ::= \psi \mid \neg\phi \mid \phi \wedge \phi \mid [\mathit{Was}: \phi] \mid [\mathit{Will}: \phi] \mid [\mathit{Sett}: \phi]$$
$$\mid [\ d\ \mathit{stit}: \phi] \mid [\ a\ \mathit{stit}: \phi] \mid [\ c\ \mathit{stit}: \phi] \mid [\vec{X} \leftarrow \vec{x}]_\lambda \phi$$

Here ψ is any element for the set of primitive events in $\mathcal{M}$ and $\vec{X} \leftarrow \vec{x}$ can be any intervention in the set $\mathcal{I}_\tau$.
The semantics of this language are the same as the semantics discussed in Section 3.3 for all the elements of the language that are also in $\mathcal{L}_G$. The sematics of the intervention operator are determined by the following rule:

$$\lambda(\mathcal{M}), m/h \models [\vec{X} \leftarrow \vec{x}]_\lambda \phi \quad \Leftrightarrow \quad \lambda(\mathcal{M}), m/h \models [d\ \mathit{stit}: \mathbf{G}\ (\vec{X} = \vec{x})] \rightarrow \phi$$

As a result of this definition, if $[\vec{X} \leftarrow \vec{x}]\vec{Y} = \vec{y}$ is true in a causal model $\mathcal{M}$ given context $\vec{u}$, the formula $[\vec{X} \leftarrow \vec{x}]_\lambda \vec{Y} = \vec{y}$ holds in all moment-history pairs of stit model $\lambda(\mathcal{M})$.

5.3 Stit and Interventions

As a result of the introduction of the $[\vec{X} \leftarrow \vec{x}]_\lambda$ operator in the previous section, it is now possible to compare expressions about interventions in a stit model $\lambda(\mathcal{M})$ to expressions in terms of the stit operators. By doing this it will become clear that every stit operator considered in this thesis has a definition that is weaker than the characterization of interventions. Consequently, when an agent intervenes on a variable in $\mathcal{M}$, it has, by all definitions of stit that are considered in this thesis, seen to it that said variable obtained its value in $\lambda(\mathcal{M})$. It is, however, possible that an agents sees to it that a variable obtains its value in $\lambda(\mathcal{M})$ without performing an action that corresponds with an intervention on that variable in $\mathcal{M}$.

5.3.1 Chellas Stit

An expression of the from $[c\ \mathit{stit}: \vec{X} = \vec{x}]$ holds at a moment-history pair m/h of a stit model $\lambda(\mathcal{M})$ if $\vec{X} = \vec{x}$ is true in m/h and all moment-history pairs Choice_m-equivalent to m/h.
When moment-history pair m/h represents intervention $\vec{X} \leftarrow \vec{x}$, this means that $h \in \mathit{Choice}_m^{\vec{X} \leftarrow \vec{x}}$. Consequently, $\lambda(\mathcal{M}), m/h \models \vec{X} = \vec{x}$. Moreover, by the definition of choice-equivalence classes, any history $h' \in \mathit{Choice}_m(h)$ choice equivalent to h is also in $\mathit{Choice}_m^{\vec{X} \leftarrow \vec{x}}$. As a result, for any moment-history pair m/h' choice equivalent to m/h, it is easy to see that $\lambda(\mathcal{M}), m/h' \models \vec{X} = \vec{x}$. $\lambda(\mathcal{M}), m/h \models [c\ \mathit{stit}: \vec{X} = \vec{x}]$ is therefore true when m/h represents intervention $\vec{X} \leftarrow \vec{x}$. Intuitively this means that when an agent intervenes on a variable in a causal model, it sees to it that that variables obtains its value according to the Chellas semantics.
Using the semantics introduced in Section 5.2 it can be deduced that the following holds in all moment-history pairs of any stit model in the image of λ.

$$[\vec{X} \leftarrow \vec{x}]_\lambda [c\ \mathit{stit}: \vec{X} = \vec{x}] \tag{5.1}$$

5.3.2 Deliberative Stit

In a moment-history pair m/h of a stit model G , $G, m/h \models [\ d\ stit:\ \phi]$ if and only if $G, m/h' \models \phi$ for each $h' \in Choice_m(h)$ and $G, m/h \not\models [Sett:\ \phi]$.
In the previous section it was shown that if in a stit model $\lambda(\mathcal{M})$, m/h represents intervention $\vec{X} \leftarrow \vec{x}$ then $\lambda(\mathcal{M}), m/h' \models \mathbf{G}\ (\vec{X} = \vec{x})$ for each $h' \in Choice_m(h)$. Since any variable can take on at least to values, there must a moment history pair m/h' corresponding to a different intervention on
$vecX$ than $\vec{X} \leftarrow \vec{x}$. In this moment-history pair $\vec{X} = \vec{x}$ does not hold and consequently $\vec{X} = \vec{x}$ is not settled at m/h.
It can be inferred that in any moment-history pair representing intervention $\vec{X} \leftarrow \vec{x}$, $[d\ stit :\ \vec{X} = \vec{x}]$ holds.

$$[\vec{X} \leftarrow \vec{x}]_\lambda [d\ stit:\ \vec{X} = \vec{x}] \tag{5.2}$$

5.4 Axioms

In (Halpern 2000), an axiomaziation for the class of recursive causal models is proposed. In this section the operator introduced in Section 5.2 is used to translate the axioms in this axiomatization to axioms in the language $\mathcal{L}_G^+$.
The axioms in this axiomatization are called **C0**, **C1**,**C2**,**C3**,**C4** and **C6**. Each axiom will be discussed in this chapter. First rule in this axiomatization, called **C0**, includes all instances off propositional tautologies. The translation of this axiom to $\mathcal{L}_G^+$ simply contains all propositional tautologies that can be expressed in $\mathcal{L}_G^+$.
Axiom **C1** is called the *equality* axiom in (Halpern 2000).

$$[\vec{Y} \leftarrow \vec{y}]X = x \Rightarrow [\vec{Y} \leftarrow \vec{y}]X \neq x' \text{ if } x, x' \in \mathcal{R}(X), x \neq x' \tag{**C1**}$$

The semantic implication of this axiom for causal models, is that in any causal model $\mathcal{M}_{\vec{Y} \leftarrow \vec{y}}$ variable X can only take on one value from the set $\mathcal{R}(X)$. Using the operator introduced in Section 5.2, the following translation of axiom **C1** can be obtained.

$$[\vec{Y} \leftarrow \vec{y}]_\lambda X = x \Rightarrow [\vec{Y} \leftarrow \vec{y}]_\lambda X \neq x' \text{ if } x, x' \in \mathcal{R}(X), x \neq x' \tag{**C1**$_\lambda$}$$

Axiom **C2** entails that given an intervention $\vec{Y} \leftarrow \vec{y}$ on a causal model $\mathcal{M}$, a variable X of that causal model must take on at least one of the values in the set $\mathcal{R}(X)$.

$$\bigvee_{x \in \mathcal{R}(X)} [\vec{Y} \leftarrow \vec{y}]X = x \tag{**C2**}$$

This can be translated easily to a statement in the language $\mathcal{L}_G^+$.

$$\bigvee_{x \in \mathcal{R}(X)} [\vec{Y} \leftarrow \vec{y}]_\lambda X = x \tag{**C2**$_\lambda$}$$

For any moment-history pair m/h of a stit model $\lambda(\mathcal{M})$, there is a vector $\vec{Y} \in \mathcal{V}$ and a vector $\vec{y} \in \mathcal{R}(\vec{Y})$ such that m/h represent causal model $\mathcal{M}_{\vec{Y} \leftarrow \vec{y}}$. This axiom can therefore be generalized to the following statement:

$$\bigvee_{x \in \mathcal{R}(X)} \lambda(\mathcal{M}), m/h \models X = x \tag{**C2**$'_\lambda$}$$

The third axiom says that given a causal model $\mathcal{M}$, an intervention $\vec{X} \leftarrow \vec{x}$ on that causal model and a context $\vec{u}$, if $Y = y$ and $W = w$ then the value of Y is not affected by intervention $W \leftarrow w$.

$$([\vec{X} \leftarrow \vec{x}]W = w \wedge [\vec{X} \leftarrow \vec{x}]Y = y) \Rightarrow [\vec{X} \leftarrow \vec{x}, W \leftarrow w]Y = y \qquad (\mathbf{C3})$$

This can be translated to the following axiom in $\mathcal{L}_G^+$.

$$([\vec{X} \leftarrow \vec{x}]_\lambda W = w \wedge [\vec{X} \leftarrow \vec{x}]_\lambda Y = y) \Rightarrow [\vec{X} \leftarrow \vec{x}, W \leftarrow w]_\lambda Y = y \qquad (\mathbf{C3}_\lambda)$$

Intuitively this axiom says that when the agent chooses a choice cell that corresponds to an intervention that sets the values of variables to their current value, the truth values of all variables will remain the same.
Axiom **C4** is as follows:

$$[X \leftarrow x, \vec{W} \leftarrow \vec{w}]X = x \qquad (\mathbf{C4})$$

Using the operator introduced in Section 5.2, the following straight forward translation of the fourth axiom can be obtained;

$$[X \leftarrow x, \vec{W} \leftarrow \vec{w}]_\lambda X = x \qquad (\mathbf{C4}_\lambda)$$

Semantically, the fourth axiom in this axiomatization entails that after intervention $\vec{X} \leftarrow \vec{x}$, any other intervention will not affect the value of X. In all moment-history pairs m/h that represent a causal model that is the result of an intervention on variable X followed by some other intervention on an arbitrary vector $\vec{W}$, the following holds;

$$[\mathit{Was} : [d \ \mathit{stit} : \mathbf{G}\ (X = x)]]$$

A translation of Axiom **C4** is therefore the following;

$$[\mathit{Was} : [d \ \mathit{stit} : \mathbf{G}\ (X = x)]] \rightarrow X = x \qquad (\mathbf{C4}_\lambda\text{'})$$

The last axiom of this axiomization is concerned with variables that affect each other.

$$(X_0 \rightsquigarrow X_1 \wedge ... \wedge X_{k-1} \rightsquigarrow X_k) \Rightarrow \neg(X_k \rightsquigarrow X_0) \qquad (\mathbf{C6})$$

In order to understand this axiom in terms of the language $\mathcal{L}_\mathcal{M}$ the semantics of $\rightsquigarrow$ must first be clarified. The expression $Y \rightsquigarrow Z$ means that the value of variable Y affects the value of variable Z. In (Halpern 2000) $Y \rightsquigarrow Z$ is defined as an abbreviation of the following formula

$$\bigvee_{\vec{X} \subset \mathcal{V}, \vec{x} \in \mathcal{R}(\vec{X}), z \neq z' \in \mathcal{R}(Z)} ([\vec{X} \leftarrow \vec{x}, Y \leftarrow y]Z = z' \wedge [\vec{X} \leftarrow \vec{x}]Z = z) \qquad (\rightsquigarrow)$$

Before considering the translation of axiom **C6** to an expression in $\mathcal{L}_G^+$, we will first consider what the $\rightsquigarrow$ operator means semantically for a stit model that is the image of a recursive causal model under λ. By literally translating the definition of $\rightsquigarrow$, the following equation is obtained;

$$Y \rightsquigarrow_\lambda Z \Leftrightarrow \bigvee_{\vec{X} \subset \mathcal{V}, \vec{x} \in \mathcal{R}(\vec{X}), z \neq z' \in \mathcal{R}(Z)} \Big([\vec{X} \leftarrow \vec{x}, Y \leftarrow y]Z = z' \wedge [\vec{X} \leftarrow \vec{x}]Z = z\Big) \qquad (5.3)$$

Axiom **C6** can easily be translated to the following axiom;

$$(X_0 \rightsquigarrow_\lambda X_1 \wedge ... \wedge X_{k-1} \rightsquigarrow_\lambda X_k) \Rightarrow \neg(X_k \rightsquigarrow_\lambda X_0) \qquad (\mathbf{C6}_\lambda)$$

Chapter 6

Causation and Applications

There are several definitions of causation that can be used to determine if events caused each other. In this chapter, a few of these definitions of causation will be introduced. All the definitions that are introduced in this chapter can be expressed in terms of the language of causal models. They will then be used to determine what caused a certain outcome in a simple example. Subsequently, it will be illustrated what happens when λ is applied to a causal model describing the same example. This will then provide a way to compare the stit semantics evaluated over the resulting stit model to the conclusions drawn by the earlier introduced definitions of causation.

6.1 Counterfactual Causation

One notion definition of causal dependency that has been discussed for several centuries if that of counterfactual dependency. This notion was first introduced by Hume (Lewis 1973). Hume stated that two events are counterfactually dependent if they both occur and if one had not occurred, the other event had also not occurred. This definition has since played an important role in the development of definitions of causation. Many definitions of causation that have since been developed are extensions of the idea of counterfactual dependency.
A more formal account of counterfactual dependency was first given by Lewis in (Lewis 1973). He used close worlds semantcis to formally define counterfactual dependency. In this paper, Lewis gives the following definition of counterfactual causation. Proposition C depends counterfactually on A if on if there are no possible A-worlds and some A-world where C holds is closer then any A-world where C does not hold.
Counterfactual causation was expressed in terms of the language of causal models in (Menzies 2017). This definition will be introduced here since the expression in terms of the language of causal models allows for reasoning over counterfactual causation in stit models in the space $\mathbb{S}_\lambda$. In (Menzies 2017), the following definition of counterfactual dependence is given:

> "A variable Y **counterfactually depends** on a variable X in a model if and only if it is actually the case that $X = x$ and $Y = y$ and there exist values $x' \neq x$ and $y' \neq y$ such that replacing the equation for X with $X = x'$ yields $Y = y'$."

This definition can be rewritten in expressions using the notation that was introduced in Chapter 3 in the following way:

Definition 6.1.1. A variable Y **counterfactually depends** on a variable X in a model $\mathcal{M}$ with context $\vec{u}$ if and only if $(\mathcal{M}, \vec{u}) \models X = x \wedge Y = y$ and there exists values $x' \in \mathcal{R}(X)$, $y' \in \mathcal{R}(Y)$ with $x' \neq x$, $y' \neq y$, and $(\mathcal{M}_{X \leftarrow x'}, \vec{u}) \models Y = y'$

6.2 Halpern and Pearl

Many definitions of causation that have been proposed since the introduction of counterfactual causation are based on the idea of counterfactual dependency that was introduced in the previous section. One of those is the definition of causation proposed in (Halpern and Pearl 2005). The definition of causation proposed in this paper is an extension of the counterfactual definition with a condition that allows for a certain set of variables to be set to fixed values while evaluating the counterfactual condition. This makes it a weaker definition, but one that is more appropriate to according to Halpern and Pearl.
In (Halpern and Pearl 2005), the following definition of causation is given:

Definition 6.2.1. $\vec{X} = \vec{x}$ is an **actual cause** of ϕ in causal model $\mathcal{M}$ with context $\vec{u}$ and set of allowable settings $\mathcal{E}$ is allowed if the following three conditions hold:

AC1. $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{x} \wedge \phi$. (That is, both $\vec{X} = \vec{x}$ and ϕ are true in the actual world.)

AC2. There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ with $\vec{X} = \vec{x}' \in \mathcal{E}$ and $\vec{W} = \vec{w}' \in \mathcal{E}$ such that if $(\mathcal{M}, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

   (a) $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}]\neg\phi$. In words, changing $(\vec{X}, \vec{W})$ from $(\vec{x}, \vec{w})$ to $(\vec{x}', \vec{w}')$ changes ϕ from true to false.

   (b) $(\mathcal{M},\ u) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\phi$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$. In words, setting any subset of variables in $\vec{W}$ to their values in $\vec{w}'$ should have no effect on ϕ, as long as $\vec{X}$ is kept at its current value $\vec{x}$, even if all the variables in an arbitrary subset of $\vec{Z}$ are set to their original values in the context $\vec{u}$.

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2. Minimality ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential for changing ϕ in AC2(a) are considered part of a cause; inessential elements are pruned.

In this definition, the variables in vector $\vec{W}$ have the role of "witness" variables while the variables in $\vec{Z}$ are considered to be part of the causal process of $\vec{X}$. The variables in $\vec{W}$ are not considered part of the causal process and are therefore kept fixed in condition AC2(a). This condition is the equivalent of Definition 6.1.1 can therefore be checked while keeping the variables in $\vec{W}$ fixed at the values that they have in model $\mathcal{M}$ under context $\vec{u}$. Halpern and Pearl state that this condition makes the definition more "permissive".

6.3 Deliberative stit in the image of λ

Applying transformation λ to stit models provides the opportunity to compare the semantics of the stit models to the semantics of causal models. This will be done later in this chapter when an example is considered. In this section, some general observations about the relation between the stit operators and definitions of causation will be made.
In Section 5.3, it has been shown that whenever an agent intervenes on a variable by making a choice in a moment-history pair of a stit model in the image of λ, it has seen to it, by both definitions of stit, that the variable that was intervened on obtained its value. In this section it will be shown that the agent has also seen to it that a variable obtains its value, when the value of that variable is counterfactually dependent on the intervention performed by the agent. This holds only for the rules of deliberative stit. When an agent performs an intervention, it has not necessarily seen to all the outcomes that are counterfactually dependent on that intervention by the rules of Chellas stit.

Theorem 2. Let $\lambda(\mathcal{M})$ be a stit model. Let X and Y be variables of $\mathcal{M}$ and let m/h a moment-history pair of $\lambda(\mathcal{M})$. If variable Y is counterfactually dependent on variable X and agent α chooses choice cell $Choice_m^{\vec{X}' \leftarrow \vec{x}'}$ in moment-history pair m/h with $X \in \vec{X}'$, then the following holds:

$$\lambda(\mathcal{M}), m/h \models [\alpha \ d \ stit : Y = y]$$

for some $y \in \mathcal{R}(Y)$

Proof. Assume that some outcome $Y = y$ is counterfactually dependent on intervention $X \leftarrow x$. Since Y is counterfactually dependent on X, $Y = y$ is true in all choice equivalent moment-history pairs of m/h. By the definition of counterfactual dependency it must be true that if the agent choose a different choice cell, $Y = y$ had not been true. Consequently, $Y = y$ cannot be true in the moment-history pairs containing m that are not choice equivalent to m/h. As a result $[\alpha \ d \ stit : Y = y]$ must holds. □

Note that the reverse of Theorem 2 is also true. Say that in some moment-history pair, $[\alpha \ d \ stit : \phi]$ holds. This means that ϕ is be true in m/h. Moreover, there is some moment-history pair of m that is not equivalent to m/h where ϕ is not true. In other words, it is possible that if α had performed a different action ϕ would not hold. ϕ is therefore counterfactually dependent on the choice of the agent.

6.4 Rock Throwing Example

The expressiveness of the counterfactual definition of causation, the definition by Halpern and Pearl, and the deliberative stit operator will be illustrated by applying them to an example. The example that will be considered is taken from (Halpern and Pearl 2005). The example concerns a situation in which two people, Billy and Suzy, simultaneously throw a rock at the bottle. Suzy's rock hits the bottle and shatters it. The other rock cannot shatter the bottle as it has already been shattered. If the Suzy's rock would not have shattered the bottle, it would have been shattered by the Billy's rock.

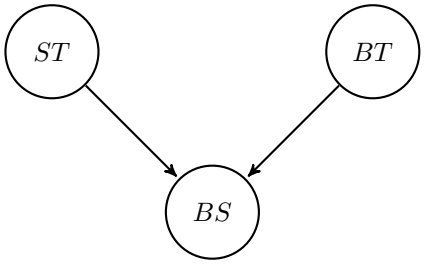

Figure 6.1: Simple Causal Model $\mathcal{M}$

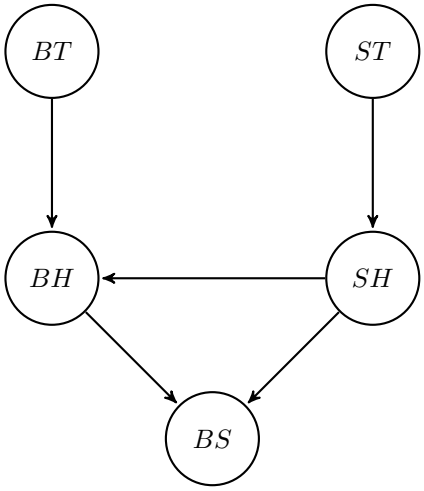

Figure 6.2: Causal Model $\mathcal{M}^S$

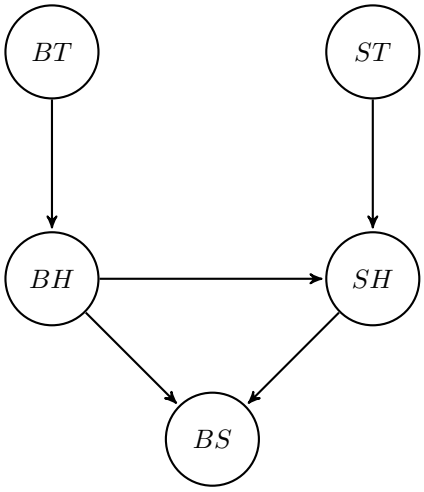

Figure 6.3: Causal Model $\mathcal{M}^B$

6.4.1 Preemtion

The example that will be used is one of preemtive causation. Specifically, this example concerns early preemption. An event is said to be a *preempted* cause of some outcome if it would have caused that outcome, had the outcome not been caused by another event (Lewis 1973). As a result, the causal process initiated by the preemted cause was interrupted before it was able to bring about the outcome.

6.4.2 Counterfactual Reasoning

As Halpern and Pearl point out, applying counter factual reasoning to this example does not provide a satisfying conclusion. If Suzy's rock had not been thrown, the bottle would still have shattered as a result of Billy's rock. According to counterfactual reasoning, the first rock did not cause the bottle to shatter. By the same reasoning it can be concluded that, according to counterfactual reasoning, the second rock being thrown did also not cause the bottle to shatter. Consequently, it can be concluded that no rock caused the bottle to shatter. It feels counter intuitive that no cause can be attributed to the bottle shattering.

6.4.3 The Halpern and Pearl Definition

It is obviously problematic that neither of the rocks is identified as the cause of the bottle shattering. In an effort to overcome this, Halpern and Pearl apply their own definition of causality to a simply causal model. This model is pictured in Figure 6.1. In this model, ST and BT are exogenous variables. BS is an endogenous variable. All variables can take on value 0 and value 1. The only structural equation of this model is $BS = \max\{ST, BT\}$.
This yields a more satisfying, but still imperfect, conclusion. Their definition applied to this causal models gives the conclusion that both the rock thrown by Billy and the rock thrown by Suzy caused the bottle to break.
By modifying the causal model representing this scenario, and applying their definition to the modified model, Halpern and Pearl eventually show that their definition of causation can determine that the first rock being thrown is the cause of the bottle shattering. They do this, however, by modifying the model in such a way that the structural equations depend on which rock reaches the bottle first. If Suzy's rock hits the bottle first, the structural equations are such that the model is as described in Figure 1.2. This model, $\mathcal{M}^S$, has two exogenous variables, BT and ST. Its set of endogenous variables consists of BH, SH, ST. All variables can take on the values 0 and 1. These variables have the following structural equations:

- $SH = ST$
- $BH = \min\{BT, (1 - SH)\}$
- $BS = \max\{SH, BH\}$

If Billy's rock hits the bottle first, the example is best described by the model in Figure 1.3. This model, $\mathcal{M}^B$ had the same signature as $\mathcal{M}^S$. It has the following structural equations:

- $BH = BT$
- $SH = \min\{ST, (1 - BH)\}$
- $BS = \max\{SH, BH\}$

Halpern and Pearl make sure that the temporal information about the order of the rocks hitting the bottle is contained in the structural equations. This seems unnatural, and as the authors admit themselves, their definition of causation is highly dependent on how the scenario is represented by the model.

6.5 Application of λ

It will now be show that by transforming the simple causal model $\mathcal{M}$ to a stit model using λ, more intuitive conclusions about the cause of the bottle shattering can be drawn without using a modified causal model that contains information about the causal of the bottle shattering.
Two different valid timings of $\mathcal{M}$ will be considered, called τ_1 and τ_2 respectively.

$$\tau_1(ST) = 1, \tau_1(BT) = 1, \tau_1(BS) = 2$$

$$\tau_2(ST) = 1, \tau_2(BT) = 2, \tau_2(BS) = 3$$

6.5.1 Model $\lambda(\mathcal{M})_1$

Let $\lambda(\mathcal{M})_1$ be the stit model built by applying transformation λ to causal model $\mathcal{M}$ with timing τ_1 and a context under which no rock is thrown. This is a multi-agent stit model in which agent $Suzy$ can intervene on ST and agent $Billy$ can intervene on BT. The set of moment of $\lambda(\mathcal{M})_1$ is denoted by $Moments^1_{\mathcal{M}}$ and has the following elements:

$$m_{[\vec{\emptyset} \leftarrow \vec{\emptyset}, \langle ST, BT \rangle]}$$

$m_{[\langle ST \leftarrow 0, BT \leftarrow 0 \rangle, BS]}$	$m_{[\langle ST \leftarrow 0, BT \leftarrow 0, BS \leftarrow 0 \rangle, \emptyset]}$	$m_{[\langle ST \leftarrow 0, BT \leftarrow 0, BS \leftarrow 1 \rangle, \emptyset]}$
$m_{[\langle ST \leftarrow 0, BT \leftarrow 1 \rangle, BS]}$	$m_{[\langle ST \leftarrow 0, BT \leftarrow 1, BS \leftarrow 0 \rangle, \emptyset]}$	$m_{[\langle ST \leftarrow 0, BT \leftarrow 1, BS \leftarrow 1 \rangle, \emptyset]}$
$m_{[\langle ST \leftarrow 1, BT \leftarrow 0 \rangle, BS]}$	$m_{[\langle ST \leftarrow 1, BT \leftarrow 0, BS \leftarrow 0 \rangle, \emptyset]}$	$m_{[\langle ST \leftarrow 1, BT \leftarrow 0, BS \leftarrow 1 \rangle, \emptyset]}$
$m_{[\langle ST \leftarrow 1, BT \leftarrow 1 \rangle, BS]}$	$m_{[\langle ST \leftarrow 1, BT \leftarrow 1, BS \leftarrow 0 \rangle, \emptyset]}$	$m_{[\langle ST \leftarrow 1, BT \leftarrow 1, BS \leftarrow 1 \rangle, \emptyset]}$

The sets $Histories^1_{\mathcal{M}}$ contains all the branches in the tree that is obtained when applying $\leq^1_{\mathcal{M}}$ to $Moments^1_{\mathcal{M}}$. And example element of $Histories^1_{\mathcal{M}}$ is the following:

$$\{m_{[\vec{\emptyset} \leftarrow \vec{\emptyset}, \langle ST, BT \rangle]}, m_{[\langle ST \leftarrow 0, BT \leftarrow 1 \rangle, BS]}, m_{[\langle ST \leftarrow 0, BT \leftarrow 1, BS \leftarrow 0 \rangle, \emptyset]}\}$$

The Choice set and valuations function of this stit model are as one would expect from the definitions in Section 4.6 and Section 4.7.
Moment $m_{[\vec{\emptyset} \leftarrow \vec{\emptyset}, \langle ST, BT \rangle]}$ is the first moment of this stit model. Agents $Suzy$ and $Billy$ both choose a choice cell at this moment. By throwing the rock, the agents choose the cells corresponding to $ST \leftarrow 1$ and $BT \leftarrow 1$ respectively. Both agents see to it that the bottle shatters according to the Deliberative and Chellas semantics. In this case, the stit logics draw a similar conclusion as the Halpern-Pearl definition applied to causal model $\mathcal{M}$.

6.5.2 Model $\lambda(\mathcal{M})_2$

Consider $\lambda(\mathcal{M})_2$, the stit model that is the result of applying λ to causal model $\mathcal{M}$ with timing τ_2 and the context under which no rocks are thrown. The set of moments of this model, $Moments^2_{\mathcal{M}}$, contains the following elements:

$m_{[\vec{\emptyset}\leftarrow\vec{\emptyset},ST]}$	$m_{[ST\leftarrow 0,BT]}$	$m_{[ST\leftarrow 0,BT]}$
$m_{[\langle ST\leftarrow 0,BT\leftarrow 0\rangle,BS]}$	$m_{[\langle ST\leftarrow 0,BT\leftarrow 0,BS\leftarrow 0\rangle,\emptyset]}$	$m_{[\langle ST\leftarrow 0,BT\leftarrow 0,BS\leftarrow 1\rangle,\emptyset]}$
$m_{[\langle ST\leftarrow 0,BT\leftarrow 1\rangle,BS]}$	$m_{[\langle ST\leftarrow 0,BT\leftarrow 1,BS\leftarrow 0\rangle,\emptyset]}$	$m_{[\langle ST\leftarrow 0,BT\leftarrow 1,BS\leftarrow 1\rangle,\emptyset]}$
$m_{[\langle ST\leftarrow 1,BT\leftarrow 0\rangle,BS]}$	$m_{[\langle ST\leftarrow 1,BT\leftarrow 0,BS\leftarrow 0\rangle,\emptyset]}$	$m_{[\langle ST\leftarrow 1,BT\leftarrow 0,BS\leftarrow 1\rangle,\emptyset]}$
$m_{[\langle ST\leftarrow 1,BT\leftarrow 1\rangle,BS]}$	$m_{[\langle ST\leftarrow 1,BT\leftarrow 1,BS\leftarrow 0\rangle,\emptyset]}$	$m_{[\langle ST\leftarrow 1,BT\leftarrow 1,BS\leftarrow 1\rangle,\emptyset]}$

The sets $Histories^2_{\mathcal{M}}$ contains all the branches in the tree that is obtained when applying $\leq^2_{\mathcal{M}}$ to $Moments^2_{\mathcal{M}}$. An example element of $Histories^2_{\mathcal{M}}$ is the following:

$$\{m_{[\vec{\emptyset}\leftarrow\vec{\emptyset},ST]}, m_{[ST\leftarrow 0,BT]}, m_{[\langle ST\leftarrow 0,BT\leftarrow 1\rangle,BS]}, m_{[\langle ST\leftarrow 0,BT\leftarrow 1,BS\leftarrow 0\rangle,\emptyset]}\}$$

This history contains one element more than the example of a history of $\lambda(\mathcal{M})_1$ that was given in this section. As a result of the greater range of τ_2, all histories of $\lambda(\mathcal{M})_2$ have one element more than histories of $\lambda(\mathcal{M})_1$. The Choice set and valuations function of this stit model are as one would expect from the definitions in Section 4.6 and Section 4.7.

In the first moment of this stit model, $m_{[\vec{\emptyset}\leftarrow\vec{\emptyset},ST]}$, agent representing *Suzy* can see to it that the bottle shatters by throwing the rock and performing intervention $ST \leftarrow 1$. If she does that, agent *Billy* cannot see to it that he shatters the bottle in the next moment. If *Suzy* does not throw the rock, *Billy* can see to it that the bottle shatters in the next moment.

These conclusions about the causal relation between the actions of the agents and the state of the bottle are more intuitive then the ones drawn by the other definitions of causation. This was possible without changing the structure of the causal model. There are two main reasons why the stit model could be used to draw the more appropriate conclusion.

Firstly, temporal information is provided by τ when applying transformation λ. This is why stit model $\lambda(\mathcal{M})$ contains more information than $\mathcal{M}$. It can therefore be used to distinguish between the rocks of Suzy and Billy without having to change the structure of the model.

Secondly, even tough the deliberative stit operator coincides with the definition of counterfactual dependency, the evaluation of the stit operator draws a more intuitive conclusion in this example then the definition of counterfactual dependency. This is because in the first moment of this model, we can evaluated what happen is Suzy does not throw, while keeping fixed that Billy did not throw. This is a result of the assumption of a context under which no rock is thrown. Consequently, deliberative stit says that Suzy sees to it that the bottle breaks. Deliberative stit and counterfactual dependence coincide, but as a result of the temporal information provided by the timing function, deliberative stit considers the effects of Suzy's throw before Billy has thrown. Counterfactual reasoning considers the effect of Suzy's throw after it is known that both Suzy and Billy will throw. It is therefor less suitable for reasoning over this example.

6.6 Conclusion

The discussion of the example in this chapter has shown that even though deliberative is similar to counterfactual dependency, it is more suitable for the evaluation of responsibility. This is because the evaluation of the deliberative stit operator considers the events that happen in chronological order. The effects of an action are evaluated only with respect to the events that have taken place before that action. Counterfactual reasoning, however, considers the effect of an action with respect to the final configuration of the causal models.
Moreover, this example has showed that in order to obtain intuitively conclusions about causation using causal models, the models had to be altered in a unnatural way while λ provides a way to add the necessary information in a more natural way by means of the timing function.

Chapter 7

Conclusion

In this thesis, stit model and causal models were introduced. The definitions and properties of both models were discussed. Subsequently, a transformation was defined that returns stit models in which each moment-history pair represents a configuration of a causal model. This transformation was used to describes a space of stit models that convey causal information. The properties of the models in this space were explored firstly by translating formulas in the causal language under transformation λ. Subsequently, their properties were explored by using the models to discuss examples from the literature in which the determination of causation is ambigous.

7.1 Observations

A few important observations were made while exploring the map from the space of causal models to the space of stit models. These will be discussed in this section. The most important observations that were made concern the properties of models in the image of λ, the expressiveness of the stit operators in these models, and how both the structure and operators contribute to expressing responsibility.

7.1.1 Properties of models

One important observation made in this thesis is that the stit models in the image of transformation λ have certain distinctive properties.
All models in this space have a symmetrical structure, in which the amount of choices available to each agent is the same for moment-history pairs in the same instant. As a result of this symmetrical structure, every configuration of a causal model $\mathcal{M}$ in represented in some moment-history pair of a stit model in $\lambda(\mathcal{M})$. Moreover, a distinctive property of models in the image of λ is that as a result of the role of interventions in these stit models, the primitive event corresponding to a choice of an agent will remain true in the rest of the history in which this choice was made.

7.1.2 Operators

Two different stit operators have been discussed in this thesis, the deliberative stit operator and the Chellas stit operator. These are similar operators, they both have the condition that the event that the agent has seen to must be true in all moment-history pairs possible after the agent has made its choice. The deliberative stit operator is more restrictive than the Chellas stit operator, since it has the extra condition that there is at least one alternative moment-history pair reachable in which this event is not true.
Intuitively, this restriction seems similar to the conditions for counterfactual dependency between the action by the agent and the event that the agent has seen to. This similarity was explored in this Section 6.3. In models in the image of transformation λ, there is indeed a connection between the deliberative stit operator and counterfactual dependency. If an agent makes a choice in a moment-history pair of such a model, the agent has seen to all events that are counterfactual dependent on the action. This comparison could only be made because a class of models was described in which *both* counterfactual dependency and stit operators could formally be expressed.
Moreover, the stit operators can be used to express interventions in terms of the stit language. Since the language of causal models is built out of primitive events and interventions, any expression from the causal language can be expressed in terms of the stit language using the transformation that was given.

7.1.3 Responsibility

The objective of this thesis is to contribute to the expression of a formal account of responsibility. In order to explore how the models described in this thesis can be used to for this objective, the models were used to represent some examples from the literature. The examples that were used concerned situations in which preemption made it difficult to determine causation.
One important finding is that in these examples the role of the timing function gives the modeller some influence about what the model will look like. The timing function has a big influence on the structure of the model. Consequently, whether an agent has seen to an outcome depends greatly on the timing function. The timing function can therefore be used to represent these example in a more intuitive way and consequently draw more appropriate conclusions on responsibility.
In order to draw appropriate conclusions about causation using the existing definitions of causation, the structure of the causal models representing the examples had to be changed. Using the stit models defined in this thesis provides a more more natural way of constructing an appropriate model.

7.2 Future research

This thesis only considered one important part of responsibility, causality, and used this to expand a branch of the logic of action. Even tough this yielded a space of models with desirable properties, the approach did not consider other aspects that are correlated to responsibility. In (Halpern and Kleiman-Weiner 2018), two other factors of responsibility are discussed. These factors are intention and blameworthiness, and Halpern expresses these factors in terms of

the causal language. Since these factors can be expressed in term of the causal language, they could be compared to some claims made about causality and the stit operators in the examples considered in this thesis. Incorporating these factors would make the discussion of responsibility more complete.
Finally, this thesis did not evaluate stit models in the image of λ for some causal model $\mathcal{M}$ with different contexts. The effects of the timing function on the conclusions that can be drawn from the model were considered. It turned out that the timing function does have some effect on the conclusions that can be drawn. The effects of the context on the conclusions could also be explored. Using a different contexts yields models with the same structure, but different valuations.

Bibliography

Beckers, Sander, and Joost Vennekens. 2018. "A principled approach to defining actual causation". *Synthese* 195, no. 2 (): 835–862. ISSN: 1573-0964. doi:10.1007/s11229-016-1247-1. https://doi.org/10.1007/s11229-016-1247-1.

Belnap, N., M. Perloff, and M. Xu. 2001. *Facing the Future: Agents and Choices in Our Indeterminist World*. Oxford University Press. ISBN: 9780195350074. https://books.google.nl/books?id=wSed02w4u-gC.

Broersen, Jan. 2014. "Responsible Intelligent Systems" [**inlang**English]. *KI - Künstliche Intelligenz* 28 (3): 209–214. doi:10.1007/s13218-014-0305-4.

Broersen, Jan, Andreas Herzig, and Nicolas Troquard. 2006. "From Coalition Logic to STIT". *Electronic Notes in Theoretical Computer Science* 157:23–35.

Cholvy, L., F. Cuppens, and C. Saurel. 1997. "Towards a Logical Formalization of Responsibility". *arXiv e-prints*.

Halpern, J. Y., and M. Kleiman-Weiner. 2018. "Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility". *arXiv e-prints* (). arXiv: 1810.05903 [cs.AI].

Halpern, Joseph Y. 2000. "Axiomatizing Causal Reasoning". *Journal of Artificial Intelligence Research* 21 (12): 317–337.

Halpern, Joseph Y, and Judea Pearl. 2005. "Causes and explanations: A structural-model approach. Part I: Causes". *British journal for the philosophy of science* 56 (4): 843–887.

Hitchcock, Christopher. 2018. "Causal Models". In *The Stanford Encyclopedia of Philosophy*, Fall 2018, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Lewis, David. 1973. "Causation". *The Journal of Philosophy* 70 (17): 556–567.

Matthias, Andreas. 2004. "The responsibility gap: Ascribing responsibility for the actions of learning automata". *Ethics and Information Technology* 6, no. 3 (): 175–183. ISSN: 1572-8439. doi:10.1007/s10676-004-3422-1. https://doi.org/10.1007/s10676-004-3422-1.

Menzies, Peter. 2017. "Counterfactual Theories of Causation". In *The Stanford Encyclopedia of Philosophy*, Winter 2017, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Moerdijk, Ieke, and Jaap van Oosten. 2018. *Sets, Models and Proofs*. Springer. ISBN: 978-3-319-92414-4.

Pearl, J. 2000. *Causality: Models, Reasoning and Inference*. Cambridge University Press. ISBN: 978-0521895606.

Segerberg, Krister, John-Jules Meyer, and Marcus Kracht. 2016. "The Logic of Action". In *The Stanford Encyclopedia of Philosophy*, Winter 2016, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.