Choice-Driven Counterfactuals

Canavotto, I.; Pacuit, E.

DOI
10.1007/s10992-021-09629-1

Publication date
2022

Document Version
Final published version

Published in
Journal of Philosophical Logic

Journal of Philosophical Logic (2022) 51:297–345
https://doi.org/10.1007/s10992-021-09629-1

Choice-Driven Counterfactuals

Ilaria Canavotto[1] · **Eric Pacuit**[2]

Received: 19 October 2020 / Accepted: 27 July 2021 / Published online: 15 October 2021

Abstract

In this paper, we investigate the semantics and logic of *choice-driven counterfactuals*, that is, of counterfactuals whose evaluation relies on auxiliary premises about how agents are expected to act, *i.e.*, about their *default choice behavior*. To do this, we merge one of the most prominent logics of agency in the philosophical literature, namely stit logic (Belnap et al. 2001; Horty 2001), with the well-known logic of counterfactuals due to Stalnaker (1968) and Lewis (1973). A key component of our semantics for counterfactuals is to distinguish between *deviant* and *non-deviant* actions at a moment, where an action available to an agent at a moment is deviant when its performance does not agree with the agent's default choice behavior at that moment. After developing and axiomatizing a stit logic with action types, instants, and deviant actions, we study the philosophical implications and logical properties of two candidate semantics for choice-driven counterfactuals, one called *rewind models* inspired by Lewis (*Nous* **13**(4), 455–476 1979) and the other called *independence models* motivated by well-known counterexamples to Lewis's proposal Slote (*Philos. Rev.* **87**(1), 3–27 1978). In the last part of the paper we consider how to evaluate choice-driven counterfactuals at moments arrived at by some agents performing a deviant action.

Keywords Counterfactuals · Stit logic · Logics of action · Logic and games · Mistakes in games

✉ Ilaria Canavotto
i.canavotto@uva.nl

Eric Pacuit
epacuit@umd.edu

[1] Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands

[2] Department of Philosophy, University of Maryland, College Park, MD, USA

1 Introduction

What would have happened if the charge nurse had not put the wrong medications on the desk? Would the intern have given them to the patient anyway? What if Alice hadn't moved out of the way? Would the thief have shot her? Would Beth's husband have picked up the kids if she hadn't? If David had bet tails, would Max have kept playing? These types of questions are asked in many situations, such as when determining responsibility, when making plans for the future, and when reasoning strategically about how our choices influence the choices of others. A common feature of these questions is that they involve *choice-driven counterfactuals*. Choice-driven counterfactuals are counterfactuals whose semantic value depends on how agents are expected to act. This means that the evaluation of a choice-driven counterfactual relies on auxiliary premises about the *default choice behavior* of the involved agents, where the default choice behavior is determined by, for instance, duties, personality, daily schedule, preferences, goals, and so on.

Our aim in this paper is to study a logic for reasoning about choice-driven counterfactuals. To do this, we merge one of the most prominent logics of agency in the philosophical literature, namely stit logic (the logic of *seeing-to-it-that*) [5, 25], with the well-known logic of counterfactuals due to Stalnaker [46] and Lewis [30].

There has been some investigation of the semantics of counterfactuals in the context of branching time [38, 49]—the theory of time that underlies stit semantics. However, these proposals do not take agency into account. In addition, although counterfactual reasoning is key to a number of applications of stit logic, such as the analysis of the notion of responsibility [2, 11, 20, 32], to our knowledge, only Xu [52] and Horty [25, Chapter 4] explicitly consider how to interpret counterfactuals in stit semantics. This paper begins to fill this important gap in the stit literature. We develop a stit logic with the resources to represent the agents' default choice behavior and show how to extend this logic with counterfactuals, highlighting some key motivating assumptions and identifying interesting logical properties of choice-driven counterfactuals.

The paper is organized as follows. In Section 2, we present the *s*tit *l*ogic with *d*eviant actions and n agents, SLD_n, that we use to study choice-driven counterfactuals. In Section 2.1, we introduce the notion of agency in branching time. In Section 2.2, we motivate a key component of our semantics for counterfactuals, namely the distinction between *deviant* and *non-deviant* actions at a moment, where an action available to an agent is deviant if it is not prescribed by the agent's default choice behavior. In Section 2.3, we present the syntax and semantics of SLD_n, and provide a sound and complete axiomatization. Section 3 extends SLD_n to include counterfactuals. In Section 3.1, we gradually introduce two candidate semantics for choice-driven counterfactuals, one called *rewind models* inspired by Lewis [31] and the other called *independence models* motivated by well-known counterexamples to Lewis's proposal [44]. The logical properties of the two semantics are studied in Section 3.2. In Section 4, we consider how to evaluate choice-driven counterfactuals at moments arrived at by some agents performing a deviant action. Finally, we conclude in Section 5 with a brief discussion of future work. All proofs are found in Appendix A and B.

2 Basic Framework

This section introduces the *s*tit *l*ogic with *d*eviant actions and n agents SLD_n that we use as a basis to study choice-driven counterfactuals. The following example, adapted from [49], illustrates the type of situation that we aim at modeling:

Example 1 There are three agents engaged in the following game: Initially, David decides whether to play with Max or Maxine and then he bets heads or tails. After David bets, the person nominated by David flips a coin. David wins if his bet matches the outcome of the coin flip and loses otherwise; Max wins just in case David loses; finally, Maxine wins no matter whether David's bet matches the outcome of the coin flip. Unknown to David, both Max and Maxine have two coins, one with heads on each side and one with tails on each side (called the *H-coin* and the *T-coin*, respectively). If Max has a chance to play, he flips the H-coin if David bets tails and the T-coin if David bets heads. If Maxine has a chance to play, she picks one of the coins to flip at random.[1] After nominating Max, David bets heads and Max flips the T-coin, so David loses.

In Example 1, after Max flips the T-coin, the counterfactual

$C1$ If David had bet tails, then he would still have lost

is intuitively true: according to the story—the reasoning goes—if David had bet tails instead of heads, Max would have flipped the H-coin, thus making David lose. In order to capture this intuition, we need a semantics that can represent the following elements:

($E1$) *The different ways in which things could go or could have gone.*
For instance, in Example 1, David bets heads but he could have bet tails, and this would have led to an alternative course of events.

($E2$) *The particular time at which an agent makes a choice.*
When we evaluate a choice-driven counterfactual, we consider what would have happened had the agents acted differently *at a particular time*. For instance, when we evaluate $C1$, we consider alternatives where David has *just* bet tails; alternatives where he has not just bet tails but did bet tails, say, two weeks ago or will bet tails six days from now are immaterial.

($E3$) *The types of action performed by the agents.*
When we evaluate a choice-driven counterfactual, we consider what would have happened had the agents performed different *types of action*. For instance, when we evaluate $C1$, we consider alternatives where David performs the action type "betting tails" instead of the action type "betting heads".

[1] The reader may wonder why we don't simply make Maxine flip a fair coin. The reason is that we will use Max and Maxine to illustrate a difference concerning the agents' default choice behavior (see p. 12), which will be important for the semantics of choice-driven counterfactuals (see p. 22). In order to illustrate this difference, it is essential that Maxine can choose between different actions, rather than only having a single choice with indeterministic outcomes available.

($E4$) *The default choice behavior of the agents.*
When we evaluate a choice-driven counterfactual, we rely on *default assumptions* about what the agents would have done had some agents acted differently. For instance, when we suppose that David bets tails in order to evaluate $C1$, we use Max's default choice behavior (*i.e.*, to select the coin that makes David lose) to conclude that he would choose the H-coin.

The semantics of stit logic has almost everything we need. Stit captures the idea that the future can unfold in different ways, and how it will actually unfold depends, in part, on what the agents decide to do. This leads to defining stit models in terms of two main components: *a branching time structure* representing the different ways things could go (as per element $E1$) and a *choice function* representing the actions available to the agents at each moment.[2] The branching time structure is sometimes supplemented with *instants*, which represent the time at which alternative moments occur (as per element $E2$); see [5]. In addition, the choice function is sometimes accompanied by a function that labels the actions available to the agents with their *types* (as per element $E3$); see, *e.g.*, [14, 27, 53]. The only missing ingredient is a representation of the agents' *default choice behavior* (element $E4$).

We propose a way to model $E4$ in Section 2.2 below, after we introduce the formal definitions of branching time structure, instant, and action-type function in Section 2.1 (readers who are familiar with these notions should feel free to skim quickly through the definitions). We then present the syntax, semantics, and an axiomatization of our stit logic with deviant actions SLD_n in Section 2.3. We will use SLD_n models to provide a semantics for choice-driven counterfactuals in Section 3.

2.1 Agency in Branching Time

A *branching time structure* is a set of moments, Mom, with a relation $<$ on Mom, where $m < m'$ means that moment m occurs before moment m'. The relation $<$ is assumed to have a treelike structure with forward branching representing the indeterminacy of the future and backward linearity representing the determinacy of the past. For technical convenience, in this paper we assume that time is *discrete*, meaning that every moment has a set of immediate successors, and that it has a *unique beginning* and *no end*. Formally:

Definition 1 (Discrete branching time structure) A *discrete branching time structure* (DBT structure) is a tuple $\langle Mom, m_0, <\rangle$, where $Mom \neq \varnothing$ is a set of moments, $m_0 \in Mom$, and $< \subseteq Mom \times Mom$ is the predecessor relation. As usual, $\leq \subseteq Mom \times Mom$ is defined as: for any $m, m' \in Mom$, $m \leq m'$ if and only if $m < m'$ or $m = m'$. The relation $<$ is assumed to satisfy the following properties: for all $m, m_1, m_2, m_3 \in Mom$,

1. *Irreflexivity*: $m \not< m$.
2. *Transitivity*: if $m_1 < m_2$ and $m_2 < m_3$, then $m_1 < m_3$.

[2] Consult [25, Chapter 2] for an overview of the key components of a stit model.

3. *Past linearity*: if $m_1 \leq m_3$ and $m_2 \leq m_3$, then either $m_1 \leq m_2$ or $m_2 \leq m_1$.
4. *Discreteness*: if $m_1 < m_2$, then there is an m_3 such that $m_1 < m_3 \leq m_2$ and there is no m_4 such that $m_1 < m_4 < m_3$.
5. *Initial moment*: $m_0 < m$.
6. *No endpoints*: there is an $m' \in Mom$ such that $m < m'$.

The standard notions used to reason about DBT structures are summarized in Table 1. Given a DBT structure $\mathcal{T} = \langle Mom, m_0, < \rangle$, each *history* $h \in Hist^{\mathcal{T}}$ represents a complete course of events. Because of forward branching, many different histories can *pass through* a single moment m (*i.e.*, m can be an element of many different histories). The set of histories passing through moment m is denoted $H_m^{\mathcal{T}}$; each $h \in H_m^{\mathcal{T}}$ represents a complete course of events that can still be realized at m. Since time is discrete with no endpoints, for each $m \in Mom$, the set of immediate successors of m, denoted $succ(m)$, is non-empty. If $h \in H_m^{\mathcal{T}}$, then $h \cap succ(m)$ is a singleton because histories are linearly ordered sets of moments. This means that there is one and only one successor of m on history h, denoted $succ_h(m)$. The condition of past linearity ensures that every non-initial moment $m \neq m_0$ has a unique predecessor, denoted $pred(m)$. An index $m/h \in Ind^{\mathcal{T}}$ represents the complete state of affairs at moment m on history h. In the context of branching time, formulas are typically evaluated at indices.

We now supplement DBT structures with instants. Intuitively, an instant is a set of moments happening at the same time.

Definition 2 (Instants) Let $\mathcal{T} = \langle Mom, m_0, < \rangle$ be a DBT structure. For any $m \in Mom$ and $n \in \mathbb{N}$, define $succ^n(m)$ recursively as follows:

1. $succ^0(m) = \{m\}$ and $succ^{n+1}(m) = \bigcup_{m' \in succ^n(m)} succ(m')$.

Table 1 Key notions related to DBT structures

Histories	•	A **history** is a maximal set of linearly ordered moments from *Mom*.
	•	$Hist^{\mathcal{T}}$ is the set of histories in $\mathcal{T}$.
	•	The elements of $Hist^{\mathcal{T}}$ are denoted with $h, h_1, h_2, \ldots, h', h'', \ldots$.
	•	History h **passes through moment** m when $m \in h$.
	•	$H_m^{\mathcal{T}} = \{h \in Hist^{\mathcal{T}} \mid m \in h\}$ is the set of histories passing through m.
	•	h_1 and h_2 are **undivided** at m iff $h_1, h_2 \in H_m^{\mathcal{T}}$ and there is an m' s.t. $m' > m$ and $m' \in h_1 \cap h_2$.
Successors	•	$succ(m) = \{m' \in Mom \mid m < m'$ and, for no $m'' \in Mom, m < m'' < m'\}$ is the set of **immediate successors** of m.
	•	If $h \in H_m^{\mathcal{T}}$, the **immediate successor of** m **on** h, denoted with $succ_h(m)$, is the unique element of $h \cap succ(m)$.
Predecessors	•	If $m \neq m_0$, $pred(m)$ is the unique **immediate predecessor** of m.
Indices	•	An **index** is a pair (m, h) such that $m \in Mom$ and $h \in H_m^{\mathcal{T}}$. We write m/h when (m, h) is an index.
	•	$Ind^{\mathcal{T}}$ is the set of indices in $\mathcal{T}$.

Then $Inst^{\mathcal{T}} = \{succ^n(m_0) \mid n \in \mathbb{N}\}$ is the set of instants over $\mathcal{T}$. We use $\mathrm{t}, \mathrm{t}_1, \mathrm{t}_2, \ldots,$ to denote elements of $Inst^{\mathcal{T}}$.

According to Definition 2, each clock tick transitions every moment in an instant to the next unique instant.[3] When $m \in \mathrm{t}$ we say that *moment* m *occurs at instant* t and when $m \in h \cap \mathrm{t}$ we say that *history* h *crosses instant* t *at moment* m. Let $\mathcal{T} = \langle Mom, m_0, < \rangle$ be a DBT structure. The fact that $<$ is discrete and rooted in m_0 ensures that:

1. $Inst^{\mathcal{T}}$ is a partition of Mom. Hence, every $m \in Mom$ occurs at one and only one instant, denoted with t_m.
2. Every history h crosses each instant t at exactly one moment, denoted with $m_{(\mathrm{t},h)}$. In what follows, we write t/h for $m_{(\mathrm{t},h)}/h$.

The above notation together with the notation introduced in Table 1 will be repeatedly used in Sections 3 and 4. In what follows, we omit the superscript $\mathcal{T}$ and simply write $Hist$, H_m, Ind, and $Inst$ when the DBT structure is clear from the context.

Turning to agency, we start by fixing sets of (names of) action types and agents:

- Let Atm be a non-empty finite set of (names of) action types. (We use a, b, c, possibly with superscripts $a', a'', \ldots,$ for elements of Atm.)
- Let $Ag = \{1, \ldots, n\}$ be the set of n agents for some number $n \in \mathbb{N}$. (We use i, j, k, possibly with superscripts $i', i'', \ldots,$ for elements of Ag.)

We think of agents as endowed with a repertoire of action types of which they can be authors. Let $Acts$ be the set of (names of) *individual actions* defined as follows:

$$Acts \subseteq Atm \times Ag$$

We write a_i when $(a, i) \in Acts$. The idea is that a_i is the action type that is instantiated whenever agent i performs an action of type a. For instance, if $a \in Atm$ is the action type "flipping a coin" and $1, 2 \in Ag$ are, respectively, David and Max, then a_1 is the action type "David flipping a coin" and a_2 is the action type "Max flipping a coin". For $i \in Ag$, let $Acts_i$ be the set of action types authored by agent i:

$$Acts_i = \{a_j \in Acts \mid j = i\}.$$

A *profile* is a function $\alpha : Ag \to Acts$ such that, for all $i \in Ag, \alpha(i) \in Acts_i$. So, a profile is any combination of actions associated with each agent. Let $Ag\text{-}Acts$ be the set of all profiles (we use Greek letters α, β, γ for elements of $Ag\text{-}Acts$). As usual, when $\alpha \in Ag\text{-}Acts$ and $I \subseteq Ag$, we will write α_I for the restriction of α to the set I, α_{-I} for $\alpha_{Ag \setminus I}$, and $\alpha(I)$ for the image of I under α.

[3] This is a convenient simplification, and is not essential for what follows. The crucial assumption is that, for $m \in Mom$, there are alternative moments occurring at the same time as m.

We make the following two key assumptions about the individual actions that are performed at a moment:

1. The action types in Atm, $Acts$, and $Ag\text{-}Acts$ represent *one-step actions*. So, in the spirit of Propositional Dynamic Logic (PDL) [22] and Coalition Logic (CL) [35], performing an action at a moment transitions to a set of *next* moments representing the different possible outcomes of the action.[4]
2. Every transition from a moment to one of its successors is brought about by a unique profile. Accordingly, we label every index m/h with the profile that brings about the transition from m to its successor on h (*i.e.*, the moment $succ_h(m)$). If index m/h is labeled with $\alpha \in Ag\text{-}Acts$, then $\alpha(i)$ represents the action type that agent $i \in Ag$ *performs* at m/h. Hence, every agent i performs one, and only one, type of action at every index m/h.

This leads us to the following definition.

Definition 3 (Action-type function) Let $\mathcal{T} = \langle Mom, m_0, < \rangle$ be a DBT structure. An action-type function over $\mathcal{T}$ is a mapping $\mathbf{act} : Ind^{\mathcal{T}} \to Ag\text{-}Acts$ that assigns to every index in $\mathcal{T}$ a profile. For any $m \in Mom$ and $i \in Ag$, let

$$Acts_i^m = \bigcup_{h \in H_m} \mathbf{act}(m/h)(i)$$

be the set of individual actions *available to agent i at m* and

$$Acts^m = \bigcup_{i \in Ag} Acts_i^m$$

be the set of individual actions *executable at m*. Then the function **act** is required to satisfy the following conditions: for all $m \in Mom$, $h_1, h_2 \in Hist$, and $i \in Ag$,

1. *No Choice Between Undivided Histories*: if h_1 and h_2 are undivided at m,[5] then $\mathbf{act}(m/h_1) = \mathbf{act}(m/h_2)$.
2. *Independence of Agents*: for all $\alpha \in Ag\text{-}Acts$, if $\alpha(j) \in Acts^m$ for all $j \in Ag$, then there is $h \in H_m$ such that $\mathbf{act}(m/h) = \alpha$.

When $|Acts_i^m| = 1$, we say that agent i has a *vacuous choice* at m.

It is not difficult to see that the set $Acts_i^m$ of actions available to agent i at moment m induces a partition on H_m: for every $h \in H_m$, the set

$$Acts_i^m(h) = \{h' \in H_m \mid \mathbf{act}(m/h')(i) = \mathbf{act}(m/h)(i)\}$$

is the cell in the partition containing h. The set $Acts_i^m(h)$ is the action token familiar in stit semantics that has been tagged with its assigned type. Note that every such

[4] We think of the assumption that the temporal ordering is discrete as a by-product of this view of actions, rather than as an assumption about the structure of time in itself.

[5] That is, $m \in h_1 \cap h_2$ and $succ_{h_1}(m) = succ_{h_2}(m)$.

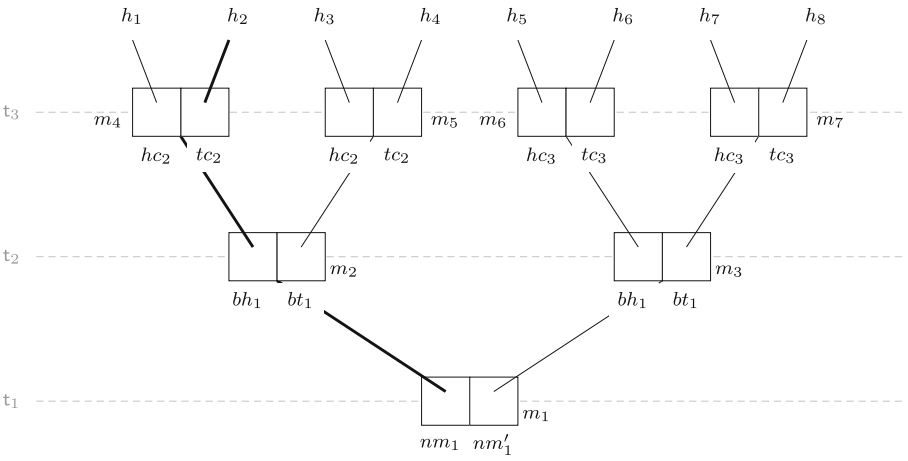

Fig. 1 Preliminary representation of Example 1

action token is assigned a unique type and different tokens are assigned different types.[6]

Conditions 1 and 2 from Definition 3 are standard requirements in stit semantics, see [25, Chapter 2]: The condition of no choice between undivided histories ensures that no individual action executable at a moment can separate histories that are undivided at that moment. The condition of independence of agents ensures that every combination of individual actions executable at a moment (one for each agent) can itself be executed at that moment.

2.2 Deviant Actions

Having introduced branching time structures, instants, and action types, the last element we need in order to provide a semantics for choice-driven counterfactuals is the notion of default choice behavior. Before presenting a formal definition, let us go back to Example 1. A DBT structure and an action-type function representing Example 1 are pictured in Fig. 1. In the figure, David is agent 1, Max is agent 2, and Maxine is agent 3. David's individual action types are nm_1 (nominate Max), nm'_1 (nominate Maxine), bt_1 (bet tails), and bh_1 (bet heads); Max's individual action types are tc_2 (flip the T-coin) and hc_2 (flip the H-coin); and Maxine's individual action types are tc_3 (flip the T-coin) and hc_3 (flip the H-coin).[7] The dashed lines represent instants, and the actual history is h_2 (the thick line).

[6]This is a common idea and can be found in, *e.g.*, [27]. It is also at the basis of the proof, presented by [16], that CL [35] can be embedded in stit.

[7]We assume that the agents who do not move at a moment m only have one available action at m—*i.e.*, the *vacuous action* (vc) that is performed on all histories passing through m. For the sake of readability, we have omitted vacuous actions from Fig. 1.

Suppose that we are at moment m_4 on history h_2 (so, David and Max have made their choices) and that we want to determine whether the counterfactual

$C1$ If David had bet tails, then he would still have lost

is true. In order to evaluate $C1$, we need to consider histories on which David performs an action of type "betting tails" *just previous to the time of* m_4 (the time of utterance). In other words, we need to consider histories on which David performs the action type bt_1 at instant t_2. Histories h_3, h_4, h_7, and h_8 all have this property. However, among these histories, we only focus our attention on those that are *most similar* to the actual history h_2. We give a full analysis of similarity in Section 3. What is important at this stage is that there is a crucial difference between h_3 and h_4.

On both histories, David bets tails at t_2 after nominating Max. Yet, after that, Max flips the H-coin on h_3 and the T-coin on h_4. The key difference is that only h_3 is consistent with Max's default choice behavior, namely that *if he has a chance to play, he flips the coin that makes David lose*. Thus, we take $C1$ to be true assuming that Max's choice matches his default choice behavior. Contrast $C1$ with the counterfactual: "If David had nominated Maxine and bet tails, then he would still have lost". Given that Maxine might well flip the T-coin, this counterfactual is false.[8]

In order to represent the default choice behavior of the agents over time, we will introduce a *deviant-action function* that identifies the *deviant actions* at each moment. An action available to an agent i at a moment m is deviant if its performance at m does not agree with agent i's default choice behavior at m—it is a *non-deviant* or *default action* otherwise. To simplify the exposition, we call an agent's default choice behavior a *choice rule*. In Example 1, "Max flips the coin that makes David lose" is a choice rule and the actions hc_2 (flipping the H-coin) and tc_2 (flipping the T-coin) are deviant actions at m_4 and m_5, respectively. The following four comments clarify the notion of choice rule.

What Choice Rules are (not). Choice rules can have various sources, including social conventions, shared standards of rationality, habits, individual preferences or goals, and, in the case of artificial agents, choice-guiding programs. Natural examples of a choice rule are the *decision rules* found in the game- and decision-theory literature, such as *expected utility maximization* or *maximin*. However, it is important to stress that some choice rules can be dictated by habits or behavior that is, on the face of it, irrational (more on this in Section 4). A final point about the interpretation of choice rules is that they should *not* be thought of as *physical or causal laws*. The key difference is that the latter laws constrain the behavior of the agents in a way that choice rules do not: while an agent who is hit on his legs by a 220 pound rolling ball cannot avoid falling, an agent who normally cheats at cards can avoid cheating.

Degrees of Deviation. It is natural to think that the notion of deviant action comes in degrees: the way that some actions deviate from the default choice behavior may be more or less important or "abnormal" than others. For simplicity, we treat all deviant

[8] What is intuitively true is "If David had nominated Maxine and bet tails, then he *might* lose".

choices equally. Everything that follows can be adapted to a graded notion of deviant action.

(In)deterministic Choice Rules. Suppose that m is a moment at which an agent i has a non-vacuous choice, and let r be a choice rule that guides the behavior of i at m. We will say that:

- r is a *deterministic* choice rule if there is only one action available to i at m that is non-deviant (the default choice behavior of i at m is *fully constrained*);
- r is an *indeterministic* choice rule if there is no action available to i at m that is deviant (the default choice behavior of i at m is *unconstrained*); and
- r is a *non-deterministic* choice rule if it is neither deterministic nor indeterministic (the default choice behavior of i at m is *partially constrained*).

Max's behavior in Example 1 is guided by a deterministic choice rule: provided that Max can play, flipping the T-coin is his only non-deviant option if David bets heads and flipping the H-coin is his only non-deviant option if David bets tails. Maxine's behavior, on the other hand, is guided by an indeterministic choice rule: if she can play, Maxine may flip either one of the two coins, no matter how David bets. Finally, an example of a non-deterministic choice rule is: "If mango, pineapple, and pear are available, then Alice picks either mango or pineapple". When all three fruits are present, this rule guides Alice's behavior only partially since picking the mango and picking the pineapple are both non-deviant. In this paper, we make the simplifying assumption that all choice rules are either deterministic or indeterministic. Excluding non-deterministic choice rules simplifies our formal definitions. Of course, this is a significant assumption since non-deterministic choice rules are ubiquitous. However, the issues concerning choice-driven counterfactuals addressed in this paper do not depend on this assumption.

Extensional Perspective on Choice Rules. Our models represent the distinction between actions that are deviant and actions that are not deviant according to an underlying set of choice rules. But we do not include a representation of the underlying choice rules themselves.[9] Using this approach, we can represent a wide variety of choice rules, including choice rules that may change over time. For example, we can easily represent the choice rule "Alice normally cheats at cards up to time t and normally respects the rules afterwards" by classifying all instances of Alice's non-cheating up to t as deviant and all instances of Alice's cheating after t as deviant. Similarly, we can represent choice rules such as "Alice is indifferent between mango and pineapple but strictly prefers watermelon over mango and pineapple": according to this rule, picking watermelon is the only non-deviant option for Alice when watermelon is available, while none of her options is deviant at moments when watermelon is not available.

We are now ready to introduce the definition of a frame for our logic SLD_n.

[9] For instance, one could make choice rules explicit using default logic as in [26]. We leave an exploration of this possibility to future work.

Definition 4 (SLD_n frame) An SLD_n frame is a tuple $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ where $\mathcal{T}$ is a DBT structure, $\mathbf{act} : Ind \to Ag\text{-}Acts$ is an action-type function over $\mathcal{T}$, and $\mathbf{dev} : Mom \to 2^{Acts}$ assigns to every moment a set of deviant individual actions. The function $\mathbf{dev}$ is required to satisfy the following conditions: for all $m \in Mom$ and $i \in Ag$,

1. *Executability of Deviant Actions*: $\mathbf{dev}(m) \subseteq Acts^m$.
2. *Availability of Non-deviant Actions*: $Acts_i^m \setminus \mathbf{dev}(m) \neq \varnothing$.
3. *(In)determinism of Choice Rules*: if $Acts_i^m \cap \mathbf{dev}(m) \neq \varnothing$, then $|Acts_i^m \setminus \mathbf{dev}(m)| = 1$.

According to condition 1, only individual actions executable at a moment can be deviant at that moment. The idea is that individual actions that cannot be performed at a moment are immaterial for the default choice behavior of the agents at that moment. According to condition 2, every agent can perform at least one non-deviant action at every moment. Given the condition of independence of agents, this means that, at every moment, there is some history on which no agent performs a deviant action. So, according to the choice rules underlying an SLD_n frame, something will always happen.[10] Finally, condition 3 captures the simplifying assumption that all choice rules are either indeterministic or deterministic. This condition ensures that, at each moment, agents can be divided into two categories: (i) agents that have no deviant actions (called *unconstrained*) and (ii) agents who have some deviant actions and only one non-deviant action (called *constrained*).[11] This distinction will play a key role in Section 3.1.

An SLD_n frame representing Example 1 is pictured in Fig. 2, where the gray cells represent the deviant actions (recall that Max's choice rule is that he flips the coin that guarantees that David bet incorrectly). In the frame, all agents are unconstrained at every moment, except for Max who is constrained at moments m_4 and m_5.

We conclude this subsection with some brief comments about extensions of the stit semantics related to the one proposed here.

The first extension that we discuss is strategic stit, see [5, Chapter 13], [25, Chapter 7], [15]. Labeling some actions as deviant at a moment can be viewed as a generalization of a strategy used in strategic stit. Given a $\mathbf{dev}$ function and an agent i, we can define a function $s_i : Mom \to 2^{Acts_i}$ as follows: for all $m \in Mom$,

$$s_i(m) = \{a_i \in Acts_i^m \mid a_i \notin \mathbf{dev}(m)\}$$

Thus defined, s_i is a partial strategy for agent i that assigns to each moment m the non-deviant actions available to i at m. It is a *partial* strategy because agent i may be unconstrained at moment m, in which case it is possible that $s_i(m) = Acts_i^m$ with $|Acts_i^m| > 1$. A similar generalization of strategic stit can be found in [33], where the authors supplement stit with a set of *rational choices* for every agent at every moment. But, as we mentioned above, choice rules may be grounded on preferences

[10] This raises an immediate question: what if a moment has been reached by some agents performing deviant actions? We discuss this issue in Section 4.

[11] Notice that, according to our terminology, agents with a vacuous choice at a moment are unconstrained at that moment, since their unique choice must be non-deviant by condition 2 in Definition 4.

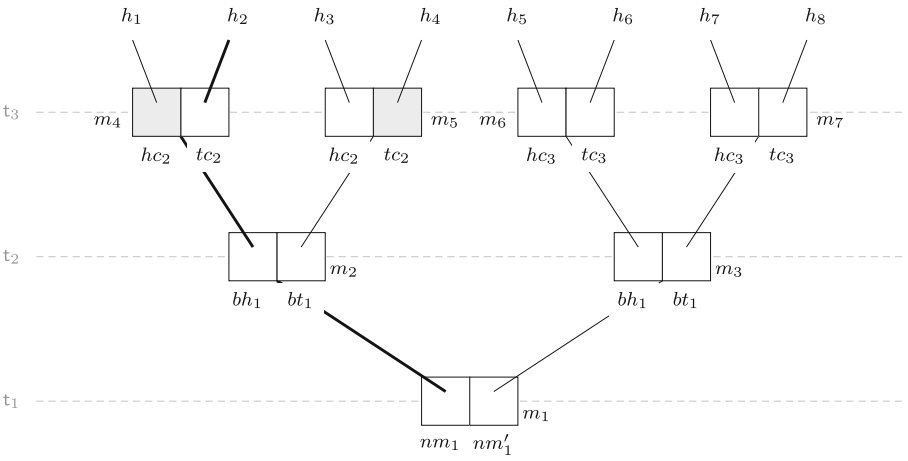

Fig. 2 SLD_n frame for Example 1

or habits that are, on the face of it, irrational. So, non-deviant choices may not coincide with rational choices. The approach that comes closest to our understanding of the **dev** function is Müller's [34, p. 199] idea of using strategic stit to "affix 'defaults' to future choices". The key difference between Müller's proposal (and, more generally, strategic stit) and our own is the role that "defaults" (or strategies) play in the semantics: in the present paper, "defaults" are introduced to contribute to the analysis of choice-driven counterfactuals rather than provide a semantics for strategic stit operators.

A second extension of stit adds epistemic operators, see, *e.g.*, [17, 23, 27, 32]. It is important to not confuse an epistemic indistinguishability relation (an equivalence relation on indices) with instants. Our interpretation of instants is that they represent "alternative presents," and *not* uncertainty of the agents. In this paper, we are interested in truth conditions for choice-driven counterfactuals, and not what such counterfactuals may express about the cognitive procedure, knowledge, and beliefs used to evaluate them.

2.3 The Logic SLD_n

Recall that $Ag = \{1, \ldots, n\}$ is a fixed set of (names of) agents and Atm is a fixed non-empty finite set of (names of) action types. In addition, let us fix a non-empty countable set $Prop$ of propositional variables (we use p, q, r, possibly with superscripts p', p'', ..., for elements of $Prop$).

Definition 5 (Syntax of SLD_n) Let $Prop$, Atm and Ag be defined as above. The set of formulas of the language of SLD_n, denoted $\mathcal{L}_{\mathsf{SLD}_n}$, is generated by the following grammar:

$$p \mid do(a_i) \mid dev(a_i) \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid \mathsf{X}\varphi \mid \mathsf{Y}\varphi$$

where $p \in Prop$ and $a_i \in Acts$.

The abbreviations for the Boolean connectives $\vee$, $\rightarrow$, $\leftrightarrow$, and the propositional constants $\bot$ and $\top$ are defined as usual. We use $\Diamond\varphi$, $\hat{\mathsf{X}}\varphi$, and $\hat{\mathsf{Y}}\varphi$ as abbreviations for $\neg\Box\neg\varphi$, $\neg\mathsf{X}\neg\varphi$, and $\neg\mathsf{Y}\neg\varphi$ respectively. Finally, we will adopt the usual rules for the elimination of parentheses.

The three modalities are standard in branching time logic: $\Box\varphi$ means "φ is settled true" or "φ is historically necessary," $\mathsf{X}\varphi$ means "φ is true at the next moment on the current history," and $\mathsf{Y}\varphi$ means "φ is true at the previous moment on the current history". The intended interpretations of the action formulas $do(a_i)$ and $dev(a_i)$ are "agent i does action a" and "action a_i is deviant", respectively. For any $\alpha \in Ag\text{-}Acts$, we define:

$$do(\alpha) := \bigwedge_{a_i \in \alpha(Ag)} do(a_i).$$

Thus, $do(\alpha)$ means "the agents do α" (*i.e.*, "for all $i \in Ag$, i performs action $\alpha(i)$").

We now define a model based on an SLD_n frame and truth for formulas from $\mathcal{L}_{\mathsf{SLD}_n}$ at an index.

Definition 6 (SLD_n model) An SLD_n model is a tuple $\mathcal{M} = \langle \mathcal{F}, \pi \rangle$, where $\mathcal{F}$ is an SLD_n frame and $\pi : Prop \rightarrow 2^{Ind}$ is a valuation function.

Definition 7 (Truth for $\mathcal{L}_{\mathsf{SLD}_n}$) Suppose $\mathcal{M}$ is an SLD_n model. Truth of a formula $\varphi \in \mathcal{L}_{\mathsf{SLD}_n}$ at an index m/h in $\mathcal{M}$, denoted $\mathcal{M}, m/h \models \varphi$, is defined recursively as follows:

$\mathcal{M}, m/h \models p$	iff	$m/h \in \pi(p)$
$\mathcal{M}, m/h \models do(a_i)$	iff	$\mathbf{act}(m/h)(i) = a_i$
$\mathcal{M}, m/h \models dev(a_i)$	iff	$a_i \in \mathbf{dev}(m)$
$\mathcal{M}, m/h \models \neg\varphi$	iff	$\mathcal{M}, m/h \not\models \varphi$
$\mathcal{M}, m/h \models (\varphi \wedge \psi)$	iff	$\mathcal{M}, m/h \models \varphi$ and $\mathcal{M}, m/h \models \psi$
$\mathcal{M}, m/h \models \mathsf{X}\varphi$	iff	$\mathcal{M}, succ_h(m)/h \models \varphi$
$\mathcal{M}, m/h \models \mathsf{Y}\varphi$	iff	$m = m_0$ or $\mathcal{M}, pred(m)/h \models \varphi$
$\mathcal{M}, m/h \models \Box\varphi$	iff	for all $h' \in H_m$, $\mathcal{M}, m/h' \models \varphi$

The notions of *validity* and *satisfiability* are standardly defined as follows: Let φ be a formula in $\mathcal{L}_{\mathsf{SLD}_n}$ and $\mathcal{M}$ an SLD_n model. Then: φ is valid in $\mathcal{M}$ just in case φ is true at all indices m/h in $\mathcal{M}$; φ is valid in the class of SLD_n models just in case φ is valid in all SLD_n models; φ is satisfiable in $\mathcal{M}$ just in case φ is true at some index m/h in $\mathcal{M}$; finally, φ is satisfiable in the class of SLD_n models just in case φ is satisfiable in some SLD_n models.

The proof of the following theorem can be found in Appendix A.

Theorem 1 *The axiom system* SLD_n*, defined by the axioms and rules in* Table 2*, is sound and complete with respect to the class of all* SLD_n *frames.*

Table 2 Axiom system SLD_n

(CPL)	Classical propositional tautologies, modus ponens		
(KD_X)	KD axioms and rules for X		
(K_Y)	K axioms and rules for Y		
($\mathsf{S5}_\Box$)	S5 axioms and rules for $\Box$		
(I) Axioms for X and Y:			
(F_X)	$\hat{\mathsf{X}}\varphi \to \mathsf{X}\varphi$	(F_Y)	$\hat{\mathsf{Y}}\varphi \to \mathsf{Y}\varphi$
(C_{XY})	$\varphi \to \mathsf{X}\hat{\mathsf{Y}}\varphi$	(C_{YX})	$\varphi \to \mathsf{Y}\hat{\mathsf{X}}\varphi$
(II) Axioms for do:			
(Act)	$\bigvee_{a_i \in Acts_i} do(a_i)$	(UH)	$(do(\alpha) \wedge \mathsf{X}\Diamond\varphi) \to \Diamond(do(\alpha) \wedge \mathsf{X}\varphi)$
(Sin)	$do(a_i) \to \neg do(b_i)$ *provided that* $a_i \neq b_i$	(IA)	$(\Diamond do(a_1) \wedge \cdots \wedge \Diamond do(a_n)) \to \Diamond do(\alpha)$ *provided that* $\alpha(1) = a_1, \ldots, \alpha(n) = a_n$
(III) Axioms for dev:			
(Ax1)	$dev(a_i) \to \Box dev(a_i)$	(Ax3)	$\bigvee_{a_i \in Acts_i}(\Diamond do(a_i) \wedge \neg dev(a_i))$
(Ax2)	$dev(a_i) \to \Diamond do(a_i)$	(Ax4)	$(\Diamond do(a_i) \wedge \Diamond do(b_i) \wedge \neg dev(a_i) \wedge dev(b_i)) \to \bigwedge_{c_i \neq a_i} (\Diamond do(c_i) \to dev(c_i))$

The axioms for do are a reformulation, in $\mathcal{L}_{\mathsf{SLD}_n}$, of the main axioms of the Dynamic Logic of Agency ($\mathcal{DLA}$) proposed by [24].[12] Axioms Act (for "Active") and Sin (for "Single") say that every agent performs one, and only one, action at every index. Axiom UH expresses no choice between undivided histories: if a group of agents performs an action that does not rule out that φ is true at the next moment, then there is some history consistent with the group action on which φ is true at the next moment. Axiom IA expresses independence of agents: if the individual actions $a_1, \ldots, a_n$ can be performed separately, then these actions can also be performed jointly.

Finally, the axioms in the last group express the fact that the **dev** function is moment-relative (axiom Ax1) and satisfies the conditions of executability of deviant actions (axiom Ax2), availability of non-deviant actions (axiom Ax3), and (in)determinism of choice rules (axiom Ax4).

3 Adding Counterfactuals

In this Section, we extend $\mathcal{L}_{\mathsf{SLD}_n}$ with formulas of the form $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ with the interpretation "if φ were true, then ψ would be true". Let $\mathcal{L}^{\Box\rightarrow}_{\mathsf{SLD}_n}$ be the full language. We

[12] It can be proved that there is a double embedding between the fragment of SLD_n without the operators dev and Y and $\mathcal{DLA}$. The reformulation of $\mathcal{DLA}$ in terms of do and X already appeared in [2].

aim at providing a semantics for $\mathcal{L}^{\Box\!\!\rightarrow}_{\mathsf{SLD}_n}$ based on SLD_n frames. Our starting point is the well-known possible world semantics for counterfactuals due to Stalnaker [46] and Lewis [30]:

($*$) A counterfactual $\varphi \Box\!\!\rightarrow \psi$ is true at a world w just in case either

(i) there is no φ-world accessible from w (the vacuous case), or
(ii) some world satisfying $\varphi \wedge \psi$ is *more similar* to w than any world satisfying $\varphi \wedge \neg\psi$.

The fundamental notion is a *relative similarity relation between possible worlds*, which [30] takes to be a *weak ordering* (a transitive relation in which ties are permitted but any two worlds are comparable) satisfying the *centering condition* (any world is more similar to itself than any other world).

There are two key questions that arise to adapt the above definition to our semantics: What should take the place of possible worlds as arguments of the relative similarity relation? What properties does the relative similarity relation satisfy? There is an extensive literature about the second question; see, *e.g.*, [6, Chapters 10-15]. While the properties we consider in this paper are not uncontroversial, our semantics for choice-driven counterfactuals takes into account some core issues from this literature. Our aim is to:

1. study the implications of these issues in our stit framework (Sections 3.1 and 3.2); and
2. explore some of the additional issues that arise when evaluating choice-driven counterfactuals after some agents don't follow their default choice behavior (Section 4).

We start with addressing the first question about the definition of relative similarity in our framework.

In the Lewis-Stalnaker semantics, possible worlds are treated as unanalyzed entities. By contrast, in our framework formulas are interpreted at a moment on a history, where the latter represents everything that happened in the past and everything that will happen in the future. From a logician's perspective, since Lewis defines relative similarity as a three-place relation on possible worlds and indices (*i.e.*, moment-history pairs) are the analogue of possible worlds in an SLD_n frame, relative similarity should be defined as a three-place relation over indices. However, when scholars in the Lewisian tradition try to put flesh on the bones of Lewis's abstract relative similarity relation, they typically think of possible worlds as evolving over time (as *histories*) and not as momentary states (as moment-history pairs).[13] This squares, too, with the analysis of Example 1 we suggested in Section 2: In order to determine the truth value of

($C1$) If David had bet tails, then he would still have lost

we consider *histories* that differ minimally from the actual one where it is true, at the time of utterance, that David bet tails and check whether, at that time, it is true that

[13] See, for instance, [31] and [6, Chapters 12-13].

David loses. From this perspective, it makes sense to introduce a *relative similarity relation between histories* (rather than indices). We will see below that, granted some additional assumptions, both perspectives can be accommodated.

Taking the more philosophical stance and following the intuitive analysis of Example 1, let us supplement SLD_n frames with a *relative similarity function*

$$\preceq\colon \mathit{Hist} \to 2^{\mathit{Hist}\times\mathit{Hist}}$$

that assigns to every history h a relative similarity relation $\preceq_h$, where for all h, h_1, h_2,

$$h_1 \preceq_h h_2$$

means "h_1 is at least as similar to h as h_2". Let a *relative similarity* SLD_n *frame* be a tuple $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq\rangle$ such that $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}\rangle$ is an SLD_n frame and $\preceq$ a relative similarity function. A *relative similarity* SLD_n *model* is a tuple $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq, \pi\rangle$ where $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq\rangle$ is a relative similarity SLD_n frame and π is a valuation function (as in Definition 6). Recall that, for any moment m, t_m is the instant to which m belongs (the time of m). When a formula is evaluated at m/h, we call t_m *the time of evaluation*. The following definition is the analogue of the Lewis-Stalnaker semantics for counterfactuals (∗):

Definition 8 (Semantics for $\varphi \boxrightarrow \psi$) Where m/h is any index from a similarity SLD_n model $\mathcal{M}$ and $\varphi, \psi \in \mathcal{L}^{\boxrightarrow}_{\mathsf{SLD}_n}$,

$\mathcal{M}, m/h \models \varphi \boxrightarrow \psi$ iff either (i) there is no $h_1 \in \mathit{Hist}$ such that $\mathcal{M}, \mathsf{t}_m/h_1 \models \varphi$
or (ii) there is $h_1 \in \mathit{Hist}$ such that $\mathcal{M}, \mathsf{t}_m/h_1 \models \varphi \wedge \psi$ and, for all $h_2 \in \mathit{Hist}$ such that $\mathcal{M}, \mathsf{t}_m/h_2 \models \varphi \wedge \neg\psi$, $h_2 \not\preceq_h h_1$

Accordingly, a counterfactual is true at an index m/h just in case the consequent is true, at the time of evaluation t_m, on all histories that differ minimally from h where the antecedent is true at t_m (if there are any histories on which the antecedent is true at t_m). We are thus assuming that the truth values of φ and ψ at indices not occurring at the time of evaluation do not affect the truth-value of $\varphi \boxrightarrow \psi$. This reflects the idea that, when we reason from a counterfactual supposition, we reason about what would happen if the supposed proposition were true *now*, see [49, p. 68]. More generally, the tense used in the antecedent and the consequent of a counterfactual is a source of indexicality: it points to a specific time (past or future) *with respect to the time of utterance*. A semantics for counterfactuals should be able to identify this specific time. Our semantics does this by first fixing the time of evaluation and then interpreting the temporal operators occurring in the antecedent and consequent.[14]

[14] Another approach would be to tag each atomic proposition with the specific time they refer to, see [42]. E.g., p_{t} means p is true at time t.

A few definitions will clarify the connection between Definition 8 and the Lewis-Stalnaker semantics (∗). For any index m/h in a similarity SLD_n model $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$, let

$$\mathsf{t}(m/h) = \{m'/h' \in \mathit{Ind} \mid \mathsf{t}_m = \mathsf{t}_{m'}\}$$

be the set of indices *accessible* from m/h. So, an index m'/h' is accessible from m/h if it occurs at the same time as m/h. Next, for any index m/h, define $\preceq_{m/h} \subseteq \mathit{Ind} \times \mathit{Ind}$ by setting, for all $m_1/h_1, m_2/h_2 \in \mathit{Ind}$:

(∗∗) $\quad m_1/h_1 \preceq_{m/h} m_2/h_2$ iff $m_1/h_1 \in \mathsf{t}(m/h)$ and $h_1 \preceq_h h_2$.

That is, m_1/h_1 is at least as similar to m/h as m_2/h_2 just in case m_1/h_1 is accessible from m/h and h_1 is at least as similar to h as h_2. The evaluation rule for $\Box\!\!\rightarrow$ in Definition 8 can then be rewritten as:

$\mathcal{M}, m/h \models \varphi \Box\!\!\rightarrow \psi$ iff either (i) there is no $m_1/h_1 \in \mathsf{t}(m/h)$ such that $\mathcal{M}, m_1/h_1 \models \varphi$,
or (ii) there is $m_1/h_1 \in \mathsf{t}(m/h)$ such that $\mathcal{M}, m_1/h_1 \models \varphi \wedge \psi$ and,
for all $m_2/h_2 \in \mathsf{t}(m/h)$ such that $\mathcal{M}, m_2/h_2 \models \varphi \wedge \neg\psi$,
$m_2/h_2 \not\preceq_{m/h} m_1/h_1$.

This is the standard evaluation rule for counterfactuals replacing possible worlds with indices. Rewriting Definition 8 in this way reveals a key assumption underlying our semantics for counterfactuals, namely that the time of evaluation does not affect the relation of relative similarity between histories: if h_1 is at least as similar to h as h_2, then this is true *no matter what time it is*. This is a substantial assumption. Contrast it with the following condition 2.3 from [49, pp. 68-69]:

> 2.3 In determining how close m_1/h_1 is to m_2/h_2 [where m_1 and m_2 occur at the same time], past closeness predominates on future closeness; that is, the portions of h_1 and h_2 not after m_1 and m_2 predominate over the rest of h_1 and h_2.
>
> This informal principle is to be intended as strongly as possible: if h_3 up to m_3 is even a little closer to h_1 up to m_1 than is h_2 up to m_2, then m_3/h_3 is closer to m_1/h_1 than m_2/h_2 is, even if h_2 after m_2 is much closer to h_1 after m_1, than h_3 after m_3. Any gain with respect to the past counts more than even the largest gain with respect to the future. [Notation adapted.]

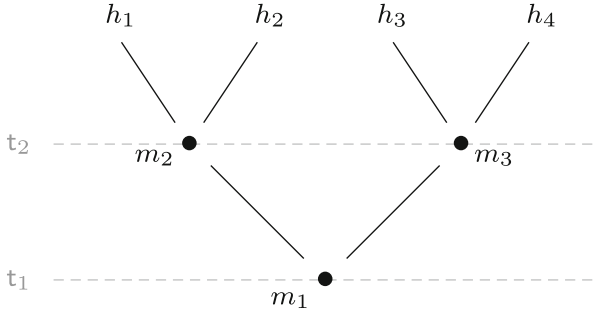

Fig. 3 Thomason and Gupta's [49] condition 2.3, an illustration

Consider the DBT structure in Fig. 3. Condition 2.3 implies that t_2/h_2 is more similar to t_2/h_1 than t_2/h_3, even if t_1/h_2 and t_1/h_3 may well be equally similar to t_1/h_1. This is excluded by our assumption $(**)$, according to which, if t_2/h_2 is more similar to t_2/h_1 than t_2/h_3, then t_1/h_2 must be more similar to t_1/h_1 than t_1/h_3. The acceptance or rejection of Thomanson and Gupta's [49] condition 2.3 influences the logic of counterfactuals. We come back to this issue in Section 3.2.

3.1 Similarity Defined

In this Section, we say more about the properties that our relative similarity relation $\preceq_h$ should satisfy.[15] We gradually introduce two candidate definitions of relative similarity in SLD_n frames. The first definition is based on Lewis's [31] criteria for determining similarity and gives rise to what we call *rewind models*. The second definition, based on well-known counterexamples to Lewis's criteria [44, p. 27, fn. 33], incorporates the idea that a notion of (in)dependence is key to a semantics of counterfactuals, giving rise to what we call *independence models*.

We start with Lewis's [31, p. 472] first criterion of similarity: "It is of the first importance to avoid big, widespread, diverse violations of law".

Lewis has in mind mainly causal or physical laws, but the notion of law in the above quote can also be understood in terms of choice rules. The suggestion is that a history h_1 is more similar to a history h than another history h_2 if fewer deviations from the agents' default choice behavior occur on h_1 than on h_2. For any history h, the *number of deviations on* h is defined as follows:

$$n\_dev(h) = \sum_{m \in h} |\{i \in Ag \mid \mathbf{act}(m/h)(i) \in \mathbf{dev}(m)\}|$$

For each history h, $n\_dev(h)$ counts, for every moment m on h, the number of agents performing a deviant action at m/h. Our first analysis of relative similarity is:

Analysis 1. For all histories h, h_1, h_2, h_1 is more similar to h than h_2 iff $n\_dev(h_1) < n\_dev(h_2)$.

Our first observation in this Section is that our definition of similarity requires additional constraints that go beyond Analysis 1. To see this, consider again Example 1 and its representation in Fig. 2. Recall that the actual history is h_2: after nominating Max, David bets heads and Max flips the T-coin, so David loses. Let L be the proposition that David loses (so, L is true at instant t_3 on h_2, h_3, h_6, h_7). Intuitively, the counterfactual $C1$ is true at m_4/h_2. The counterfactual $C1$ is expressed by the following formula of $\mathcal{L}_{\mathsf{SLD}_n}$:

$(F1)$ $\mathsf{Y}do(bt_1) \Box\!\!\rightarrow L$ ("If David had bet tails, then he would still have lost").

[15] As [6, p. 196] notes: "Lewis's theory evidently needs to be based [...] on a similarity relation that is constrained somehow—it must say that $A \Box\!\!\rightarrow C$ is true just in case C is true at the A-worlds that are most like the actual world in *such and such respects*. The philosophical task is to work out *what* respects of similarity will enable the theory to square with our intuitions and usage".

It is not hard to see that Definition 8 and Analysis 1 would evaluate $F1$ as false. The histories on which $\mathsf{Y}do(bt_1)$ is true at the time of evaluation $\mathsf{t}_{m_4} = \mathsf{t}_3$ are h_3, h_4, h_7, and h_8. Among these histories, the ones with the fewest number of deviations are h_3, h_7, and h_8 (in fact, no deviant action is performed on these histories). So, according to Analysis 1, h_3, h_7, and h_8 are the most similar histories to h_2 on which $\mathsf{Y}do(bt_1)$ is true at t_3. But $\neg L$ rather than L is true on h_8 at t_3. So, if we compare histories only in terms of the number of deviations as in Analysis 1, then $F1$ turns out to be false at m_4/h_2. The problem with Analysis 1 is that it ignores the fact that a "small miracle" [31, p. 478] (or a "surgical intervention" [36, p. 239]) at m_4/h_2 suffices to reach h_3 from h_2, while a substantial change in the past is needed to reach h_7 and h_8. This suggests that *the greater past overlap between h_3 and h_2 is more important than the fewer number of deviations on h_7 and h_8.*

Given the condition of past linearity, the *past overlap between two histories h_1 and h_2* is their intersection:[16]

$$past\_ov(h_1, h_2) = h_1 \cap h_2$$

This leads to a straightforward modification of Analysis 1:

Analysis 2. For all histories h, h_1, h_2, h_1 is more similar to h than h_2 iff
either $past\_ov(h, h_1) \supset past\_ov(h, h_2)$,
or $past\_ov(h, h_1) = past\_ov(h, h_2)$ and $n\_dev(h_1) < n\_dev(h_2)$.

Remark 1 The criterion of past overlap is the second criterion for determining similarity between histories proposed by [31]. There are well-known criticisms of this criterion: Suppose you left your jacket on a chair in a café. Consider the counterfactual "If my jacket had been stolen, then it would have been stolen right before I left". Since the histories on which your jacket has been stolen one moment ago have the greatest past overlap with the current history, the past overlap criterion implies that this counterfactual is true. This is clearly a counterintuitive consequence of past overlap. However, this issue arises when evaluating a counterfactual whose antecedent includes an arbitrary past operator. The closest we can come to express this counterfactual is "If my jacket had been stolen n moments ago, then it would have been stolen one moment ago," which is clearly false when $n > 1$. In this paper we assume the Lewisian analysis and leave a full discussion of this problem for future work. In doing this, we follow previous work on the semantics of counterfactuals in the context of branching time [38, 52], where a relative similarity relation between histories is defined in terms of the past overlap criterion. Unlike in the present paper, these papers do not consider any other criterion of similarity.

Analysis 2 delivers the correct evaluation of $F1$ at m_4/h_2: Histories h_3 and h_4 are more similar to h_2 than h_7 and h_8, because their past overlap with h_2 is greater. In turn, history h_3 is more similar to h_2 than h_4 because there are fewer deviations on

[16]The condition of past linearity ensures that $h_1 \cap h_2$ is an initial segment of both h_1 and h_2. This is why it makes sense to call it their *past* overlap.

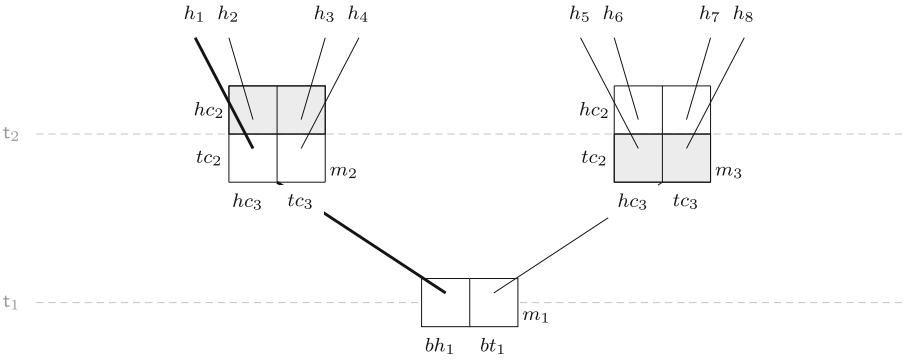

Fig. 4 Max and Maxine flip the coin at the same time

h_3 than on h_4. Since David loses at t_3 on history h_3, $F1$ is true at m_4/h_2. However, there are still problems with Analysis 2, as illustrated by the following example:

Example 2 Everything is as in Example 1 except that David does not initially nominate Max or Maxine. Instead both Max and Maxine flip a coin after David bets. David wins only if both Max's and Maxine's coins land on the side he bets. Suppose that after David bets heads, Max flips the T-coin (as prescribed by his choice-rule) and Maxine happens to flip the H-coin. So David loses.

An SLD_n frame representing Example 2 is depicted in Fig. 4, where the labels and shadings are read as in Fig. 2 on page 12 and the proposition L that David loses is true at instant t_2 on all histories except for h_2 and h_8. The actual history is h_1 (the thick line). Consider the following counterfactual:

$(F2)$ $\quad do(hc_2) \boxrightarrow \neg L$ ("If Max flipped the H-coin, then David would have won").

Intuitively, $F2$ is true at m_2/h_1. But Analysis 2 and Definition 8 do not vindicate this judgement. The histories on which Max flips the H-coin at $\mathsf{t}_{m_2} = \mathsf{t}_2$ are h_2, h_3, h_6, and h_7. Histories h_2 and h_3 have a greater past overlap with h_1 than h_6 and h_7, so the latter two histories can be discarded. In turn, since the number of deviations on h_2 is the same as the number of deviations on h_3, h_2 and h_3 are equally similar to h_1. Yet, L rather than $\neg L$ is true on h_3 at t_2. Given Definition 8, it follows that David *might* win—a weaker conclusion than the desired one. The problem is that, even though h_2 and h_3 have the same past overlap with h_1 as well as the same number of deviations, more agents need to change their actions to reach h_3 than h_2 (in this sense the change required to reach h_3 is not *minimal*). This suggests that *the smaller change making h_2 branch off from h_1 is more important than the equal number of deviations on h_2 and h_3.*[17]

[17]The importance of fixing the actions of as many agents as possible when evaluating a counterfactual in a stit model is already emphasized by Horty [25, Chapter 4], who uses this criterion to define a selection function that picks, for every index m/h, agent i, and action (token) K available to i at m, the most similar histories to h where i performs K. Since he is only interested in counterfactuals of form "if

Given two histories h_1 and h_2, say that h_1 and h_2 *divide at moment* m if m is the last moment they share, *i.e.*, $m \in h_1 \cap h_2$ and $succ_{h_1}(m) \neq succ_{h_2}(m)$. When h_1 and h_2 divide at moment m, let the *number of agents separating* h_1 *and* h_2 be defined as follows:

$$n\_sep(h_1, h_2) = |\{i \in Ag \mid \mathbf{act}(m/h_1)(i) \neq \mathbf{act}(m/h_2)(i)\}|$$

Then, $n\_sep(h_1, h_2)$ counts the number of agents that, by performing different actions on h_1 and h_2 at moment m, make h_1 and h_2 divide at m.[18] When h_1 and h_2 never divide (*i.e.*, $h_1 = h_2$), let $n\_sep(h_1, h_2) = 0$. Putting everything together, we have our first definition of similarity.

Definition 9 (Rewind similarity function) Let $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ be an SLD_n frame. Define

$$\prec^R : Hist \to 2^{Hist \times Hist}$$

by setting: for all $h, h_1, h_2 \in Hist$: $h_1 \prec^R_h h_2$ iff:

$past\_ov(h, h_1) \supset past\_ov(h, h_2)$, or
$past\_ov(h, h_1) = past\_ov(h, h_2)$ and $n\_sep(h, h_1) < n\_sep(h, h_2)$, or
$past\_ov(h, h_1) = past\_ov(h, h_2)$ and $n\_sep(h, h_1) = n\_sep(h, h_2)$ and $n\_dev(h_1) < n\_dev(h_2)$.

Define $\preceq^R_h$ as follows: for all $h_1, h_2 \in Hist$:

$$h_1 \preceq^R_h h_2 \text{ iff either } h_1 \prec^R_h h_2 \text{ or } (h_1 \not\prec^R_h h_2 \text{ and } h_2 \not\prec^R_h h_1).$$

We will call *rewind model* any similarity model $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq^R, \pi \rangle$, where $\preceq^R$ is defined as in Definition 9.

Definition 9 encodes a substantial assumption about how we let a scenario unfold under the supposition that the antecedent of a counterfactual is true. To see this, let us go back to our initial Example 1 (cf. also Fig. 2, p. 12), but suppose that the actual history is h_6 instead of h_2: After nominating Maxine, David bets heads and Maxine happens to flip the T-coin, so David loses. What if David had bet tails? Would he have won? There are two ways to answer this question.

(1) *Rewind History*: When we suppose that David bet differently, we *rewind* the course of events to the moment when David bets (m_3), intervene on his choice, and then let the future unfold according to the agents' default choice behavior. Since there is no choice rule constraining Maxine's flip, we only conclude that David *might* win. This is the conclusion we reach by applying Definition 9,

agent i performed (now) a different action, then φ would be true," [25] does not consider other criteria of similarity.

[18] Notice that, by the condition of past linearity, if two histories h_1 and h_2 divide at a moment, then they divide at a unique moment, so $n\_sep(h_1, h_2)$ is well defined.

according to which h_3 and h_4 are equally similar to h_2. In fact, together with Definition 8, Definition 9 encodes the following Lewisian procedure:

> [T]ake the counterfactual present, avoiding gratuitous difference from the actual present; graft it smoothly onto the actual past; let the situation evolve according to the actual laws; and see what happens. [31, p. 463]

(2) *Assume Independence*: When we suppose that David bet differently, we rewind the course of events to the moment when David bets (m_3), intervene on his choice, *leave all events that are independent of it as they actually are*, and *then* let the future unfold according to the agents' default choice behavior. Doing otherwise "would seem to be positing some strange causal influence" [49, p. 83]. Since there is no choice rule according to which Maxine's choice depends on David's bet, we conclude that, if David had bet differently, then he would have won.

To make the reasoning in (2) precise, we need to identify all the events that are independent of David's choice. In stit, we can think of events as actions performed by agents (possibly treating Nature as an agent). This allows us to use our distinction between constrained and unconstrained agents to capture the reasoning in (2): the unconstrained agents whose default choice behavior is not constrained by a choice rule at a moment are precisely those whose actions at that moment are independent of the actions performed at previous moments (*e.g.* David betting).[19]

To account for the Assume Independence intuition, we supplement Definition 9 with a further requirement on *unconstrained agents*. Recall that an agent i is unconstrained at a moment m when none of the actions available to her at m is deviant (cf. Section 2.2). The *set of agents unconstrained at moment m* is thus defined as:

$$\underline{Ag}(m) = \{i \in Ag \mid Acts_i^m \cap \mathbf{dev}(m) = \varnothing\}$$

Given an index m/h, define the *set of actions performed by unconstrained agents at* m/h as:

$$\underline{\mathbf{act}}(m/h) = \{\mathbf{act}(m/h)(i) \mid i \in \underline{Ag}(m)\}$$

Then the number of independent events for any histories h_1 and h_2 is defined as:

$$n\_indep(h_1, h_2) = \sum_{\mathrm{t} \in Inst} |\underline{\mathbf{act}}(\mathrm{t}/h_1) \cap \underline{\mathbf{act}}(\mathrm{t}/h_2)|$$

[19] To account for the reasoning in (2) in the context of branching time, Thomason and Gupta [49] impose constraints of "causal coherence" on their models. Yet, they acknowledge that this move adds a substantial layer of complexity to their theory. With a similar aim but in the context of branching space-time, Placek and Müller [38] define "independence" as space-like separation. Yet, they acknowledge that this kind of independence is hardly realized in everyday situations like the betting scenarios of our examples. The possibility of distinguishing constrained and unconstrained agents provides us with a convenient way to get around these difficulties.

Thus, *n_indep* counts, for every instant t, the *number of agents unconstrained at* t *on both* h_1 *and* h_2 *that act in the same way on these histories*.[20] Let us illustrate the previous definitions with Fig. 2. Assume that the vacuous choices of agent $i \in \{1, 2, 3\}$ are all labeled with vc_i. We then have the following:

- $\underline{Ag}(m_k) = \{1, 2, 3\}$ for $k \in \{1, 2, 3, 6, 7\}$ and $\underline{Ag}(m_j) = \{1, 3\}$ for $j \in \{4, 5\}$;
- $\underline{\mathbf{act}}(\mathrm{t}_1/h_1) \cap \underline{\mathbf{act}}(\mathrm{t}_1/h_5) = \{vc_2, vc_3\}$,
  $\underline{\mathbf{act}}(\mathrm{t}_2/h_1) \cap \underline{\mathbf{act}}(\mathrm{t}_2/h_5) = \{bh_1, vc_2, vc_3\}$,
  $\underline{\mathbf{act}}(\mathrm{t}_3/h_1) \cap \underline{\mathbf{act}}(\mathrm{t}_3/h_5) = \{vc_1\}$,
  so $n\_indep(h_1, h_5) = n\_indep(h_5, h_1) = 6$;
- $\underline{\mathbf{act}}(\mathrm{t}_1/h_5) \cap \underline{\mathbf{act}}(\mathrm{t}_1/h_7) = \{nm'_1, vc_2, vc_3\}$,
  $\underline{\mathbf{act}}(\mathrm{t}_2/h_5) \cap \underline{\mathbf{act}}(\mathrm{t}_2/h_7) = \{vc_2, vc_3\}$,
  $\underline{\mathbf{act}}(\mathrm{t}_3/h_5) \cap \underline{\mathbf{act}}(\mathrm{t}_3/h_7) = \{vc_1, vc_2, hc_3\}$,
  so $n\_indep(h_5, h_7) = n\_indep(h_7, h_5) = 8$.

Our second definition of similarity refines our first definition by incorporating the assumption of independence discussed in item (2) above.

Definition 10 (Independence similarity function) Let $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ be an SLD_n frame. Define

$$\prec^I : Hist \to 2^{Hist \times Hist}$$

by setting: for all $h, h_1, h_2 \in Hist$: $h_1 \prec^I_h h_2$ iff either one of the first two conditions in Definition 9 is satisfied or one of the following holds:

$$\begin{aligned} past\_ov(h, h_1) = past\_ov(h, h_2) &\text{ and } n\_sep(h, h_1) = n\_sep(h, h_2) \\ &\text{and } n\_indep(h, h_1) > n\_indep(h, h_2), \text{ or} \\ past\_ov(h, h_1) = past\_ov(h, h_2) &\text{ and } n\_sep(h, h_1) = n\_sep(h, h_2) \\ &\text{and } n\_indep(h, h_1) = n\_indep(h, h_2) \text{ and } n\_dev(h_1) < n\_dev(h_2) \end{aligned}$$

Define $\preceq^I_h$ as follows: for all $h_1, h_2 \in Hist$,

$$h_1 \preceq^I_h h_2 \text{ iff either } h_1 \prec^I_h h_2 \text{ or } (h_1 \not\prec^I_h h_2 \text{ and } h_2 \not\prec^I_h h_1).$$

We will call *independence model* any similarity model $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq^I, \pi \rangle$, where $\preceq^I$ is defined as in Definition 10. In the following, we will use $\prec$ for elements of $\{\prec^R, \prec^I\}$ and $\preceq$ for elements of $\{\preceq^R, \preceq^I\}$.

Definition 10 delivers the correct analysis of Example 2: although h_2 and h_3 overlap the same initial segment of h_1, at m_2 both David and Maxine act in the same way on h_2 and h_1, while Maxine changes her behavior on h_3. Hence, h_2 is more similar to h_1 than h_3. Since $\neg L$ is true on h_2 at t_2, it follows that $F2$ is true at m_2/h_1.[21]

[20] The reason why *n_indep* is defined over all instants rather than a single instant or a set of relevant instants is that our relative similarity relation compares histories "globally" (see the discussion on pp. 17-18)

[21] Note that this analysis essentially relies on the assumption that Maxine has *two* choices: she can pick the H-coin or pick the T-coin. If Maxine tossed a fair coin instead of choosing between the H-coin and the T-coin, the example would be different since Maxine would have a *single* choice with indeterministic outcomes instead of two choices with deterministic outcomes. So, unless the coin itself was modeled as an unconstrained agent (*i.e.*, treat nature as an agent), our analysis would be different.

3.2 Logical Properties

The following are some immediate consequences of Definitions 9 and 10.

Proposition 1 *Suppose that* $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$ *is either a rewind model or an independence model. For any history* h*, the relative similarity relation* $\preceq_h$ *is a centered weak ordering. That is,* $\preceq_h$ *satisfies the following conditions: for any* $h', h_1, h_2, h_3 \in Hist$*,*

1. *Transitivity: if* $h_1 \preceq_h h_2$ *and* $h_2 \preceq_h h_3$*, then* $h_1 \preceq_h h_3$*.*
2. *Linearity: either* $h_1 \preceq_h h_2$ *or* $h_2 \preceq_h h_1$*.*
3. *Centering: if* $h' \preceq_h h$*, then* $h' = h$*.*

Recall that, for any index m/h from a similarity SLD_n model, the set of indices accessible from m/h is $\mathsf{t}(m/h) = \{m'/h' \in Ind \mid \mathsf{t}_m = \mathsf{t}_{m'}\}$. The following is a straightforward corollary of Proposition 1:

Corollary 1 *Suppose that* $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$ *is either a rewind model or an independence model. For any index* m/h*, the relation* $\preceq_{m/h} \subseteq Ind \times Ind$ *defined by setting: for all* $m_1/h_1, m_2/h_2 \in Ind$*,*

$(**)$ $\quad m_1/h_1 \preceq_{m/h} m_2/h_2$ *iff* $m_1/h_1 \in \mathsf{t}(m/h)$ *and* $h_1 \preceq_h h_2$*,*

is a centered weak ordering satisfying the following: for all $m_1/h_1, m_2/h_2 \in Ind$*,*

Priority: if $m_1/h_1 \in \mathsf{t}(m/h)$ *and* $m_2/h_2 \notin \mathsf{t}(m/h)$*, then* $m_2/h_2 \not\preceq_{m/h} m_1/h_1$

Definition 8 together with Corollary 1 show that our semantics for the counterfactual matches the Lewisian semantics with possible worlds replaced with indices. Proposition 2 is then a well-known consequence of Corollary 1.

Proposition 2 *The following axioms and rule are valid and truth preserving in any rewind model (resp. independence model):*[22]

$(\mathsf{K}_{\Box\!\!\rightarrow})$	$(\varphi \Box\!\!\rightarrow (\psi_1 \rightarrow \psi_2)) \rightarrow ((\varphi \Box\!\!\rightarrow \psi_1) \rightarrow (\varphi \Box\!\!\rightarrow \psi_2))$
(Suc)	$\varphi \Box\!\!\rightarrow \varphi$
(Inc)	$(\neg\varphi \Box\!\!\rightarrow \varphi) \rightarrow (\psi \Box\!\!\rightarrow \varphi)$
(Cen)	$\varphi \rightarrow (\psi \leftrightarrow (\varphi \Box\!\!\rightarrow \psi))$
(Cond)	$((\varphi_1 \wedge \varphi_2) \Box\!\!\rightarrow \psi) \rightarrow (\varphi_1 \Box\!\!\rightarrow (\varphi_2 \rightarrow \psi))$
(RMon)	$\neg(\varphi_1 \Box\!\!\rightarrow \neg\varphi_2) \wedge (\varphi_1 \Box\!\!\rightarrow \chi) \rightarrow ((\varphi_1 \wedge \varphi_2) \Box\!\!\rightarrow \chi)$
$(\mathsf{RN}_{\Box\!\!\rightarrow})$	*From* ψ *infer* $\varphi \Box\!\!\rightarrow \psi$

[22] Suc stands for "Success," Inc for "Inclusion" (as it says that the closest indices satisfying a counterfactual antecedent—if any—are accessible), Cen stands for "Centering," Cond for "Conditionalization," and RMon for "Rational Monotonicity".

More interestingly, the principles in the next proposition reflect the interaction between counterfactuals and temporal modalities.

Proposition 3 *The following principles are valid in any rewind model (resp. independence model).*

$$(\mathsf{Dis}_{\mathsf{X}}) \quad \mathsf{X}(\varphi \boxrightarrow \psi) \leftrightarrow (\mathsf{X}\varphi \boxrightarrow \mathsf{X}\psi) \qquad (\mathsf{Dis}_{\mathsf{Y}}) \quad \mathsf{Y}(\varphi \boxrightarrow \psi) \leftrightarrow (\mathsf{Y}\varphi \boxrightarrow \mathsf{Y}\psi)$$
$$(\mathsf{Cen1}) \quad \Diamond\varphi \rightarrow (\Diamond\psi \rightarrow \Box(\varphi \boxrightarrow \Diamond\psi)) \qquad (\mathsf{Cen2}) \quad \Diamond\varphi \rightarrow ((\varphi \boxrightarrow \Box\psi) \rightarrow \Box\psi)$$

Proof See Appendix B. □

Corollary 2 *The following principles are theorems of the axiom system obtained by extending* SLD_n *with the principles in Proposition 2,* Cen1 *and* Cen2*:*

1. $\Diamond\varphi \rightarrow (\Box\psi \leftrightarrow \Box(\varphi \boxrightarrow \Box\psi))$ 2. $\Diamond\varphi \rightarrow ((\varphi \boxrightarrow \Box\psi) \rightarrow \Box(\varphi \boxrightarrow \psi))$

Proof Straightforward given Cen1, Cen2, and the fact that $\Box$ is an S5 modality. □

The validity of the distribution principles $\mathsf{Dis}_{\mathsf{X}}$ and $\mathsf{Dis}_{\mathsf{Y}}$ depends on the assumption that the time of evaluation does not affect the relation of relative similarity between histories. In fact, since the most similar histories to a history h *up to the present time* t are the same as the most similar histories to h *up to one instant after* t, the most similar histories to h on which $\mathsf{X}\varphi$ is true *at* t must be the same as the most similar histories to h on which φ is true *one instant after* t (similarly for $\mathsf{Y}\varphi$).

Interestingly, the condition 2.3 from [49] (see p. 18) makes it possible to find counterexamples to $\mathsf{Dis}_{\mathsf{X}}$ and $\mathsf{Dis}_{\mathsf{Y}}$. To see this, let us go back to Fig. 3. Recall that, according to condition 2.3, t_2/h_2 is more similar to t_2/h_1 than t_2/h_3. Assume that t_1/h_2 and t_1/h_3 are equally similar to t_1/h_1 and that p is true only at t_2/h_2 and t_2/h_3 while q is true only at t_2/h_2. Since q is true at the most similar index to t_2/h_1 at which p is true (*i.e.*, t_2/h_2), $p \boxrightarrow q$ is true at t_2/h_1, and so $\mathsf{X}(p \boxrightarrow q)$ is true at t_1/h_1. On the other hand, since $\neg\mathsf{X}q$ is true at one of the most similar indices to t_1/h_1 at which $\mathsf{X}p$ is true (*i.e.*, t_1/h_3), $\mathsf{X}p \boxrightarrow \mathsf{X}q$ is false at t_1/h_1.

Thomason and Gupta [49, pp. 70-71] rely on a variant of Example 1 to support the claim that $\mathsf{Dis}_{\mathsf{X}}$ and $\mathsf{Dis}_{\mathsf{Y}}$ should not come out as logical validities. In their version of the example, Max and David are the only agents, the game starts with David's bet (at t_2 in Fig. 2) and ends after Max flips either the T-coin or the H-coin. So we can depict their example as in Fig. 2 ignoring histories h_5, h_6, h_7, and h_8 and moments occurring before time t_2. As in Example 1, Max flips the coin that guarantees that David loses. In addition, the actual history is h_2: David bets heads and Max flips the T-coin. Now, let L' be the proposition "David loses at time t_3" (so, L' is true at all moments on histories h_2 and h_3). According to [49], the counterfactual

(*A*) $do(bt_1) \boxrightarrow L'$ ("If David bets tails, he would lose at t_3")

is *intuitively true* at t_2/h_2, *i.e.*, *at the beginning of the game* on the actual history. Hence, $\mathsf{Y}(do(bt_1) \rightarrow L')$ is true at t_3/h_2. On the other hand, the authors take the counterfactual

(B) $\quad \mathsf{Y}do(bt_1) \boxrightarrow \mathsf{Y}L'$ ("If David had bet tails, he would have lost at t_3")

to be *intuitively false* at t_3/h_2, *i.e.*, *at the end of the game* on the actual history. If this is correct, then the implication $\mathsf{Y}(do(bt_1) \boxrightarrow L') \rightarrow (\mathsf{Y}do(bt_1) \boxrightarrow \mathsf{Y}L')$ is false at t_3/h_2, that is, the principle $\mathsf{Dis_Y}$ is *not* intuitively valid.[23]

We disagree with Thomanson's and Gupta's judgement about B. Given Max's choice rule, at the end of the game it would be perfectly natural to explain to David: "Well, if you had bet tails, you would still have lost". We think that the problem stems from a confusion between the *time of evaluation* and the *time to which the antecedent of a counterfactual refers*. In discussing the present example, Thomason and Gupta seem to take it that, in reasoning from a counterfactual supposition, we hold fixed as many past facts as possible up to the time of evaluation (t_2 in the case of A and t_3 in the case of B). But, as most scholars think (cf. [6, Chapter 12]), what we intuitively do is rather to hold fixed as many past facts as possible up to the time to which the antecedent refers (t_2 for both A and B).[24] It then makes sense that relative similarity between histories is not affected by the time of evaluation: what is important is just that the longer a history h' overlaps another history h, the more similar h' is to h.

Turning to $\mathsf{Cen1}$ and $\mathsf{Cen2}$, the validity of these principles follows from the priority of the criterion of past overlap: if φ *can* be true at a *moment*, then supposing that φ is true does not require shifting to a different moment. (Compare the reasoning behind the validity of Cen: if φ *is* true at an *index*, then supposing that φ is true does not require moving to a different index).

Items 1 and 2 in Corollary 2 highlight an interesting interaction between counterfactuals and historical necessity. In particular, item 2, which we discuss below, can be viewed as a principle of "exportation" of $\Box$ from $\boxrightarrow$.

The validities we have considered so far do not depend on whether we work with rewind models or with independence models. The next Proposition 2 involves a formula that distinguishes the two classes of models.

[23] Observe that Thomason and Gupta's [49] condition 2.3 does not exclude the possibility of defining a similarity relation between the indices from Fig. 2 such that t_2/h_3 is the most similar index to t_2/h_2 where $do(bt_1)$ is true and t_3/h_4 is the most similar index to t_3/h_2 where $\mathsf{Y}do(bt_1)$ is true. Given such a similarity relation, A turns out to be true at t_2/h_2 while B turns out to be false at t_3/h_2, in accordance with the authors' intuitive judgement. Our property $(**)$ does not allow us to define a similarity relation of this sort: according to it, t_2/h_3 is the most similar index to t_2/h_2 where $do(bt_1)$ is true *if and only if* t_3/h_3 is the most similar index to t_3/h_2 where $\mathsf{Y}do(bt_1)$ is true.

[24] It is worth noting that, if we kept fixed as many past facts as possible up to the time of evaluation, B would be false, no matter whether Max flips the T-coin by chance or because his default choice behavior is to make David lose. Yet, intuitively, we judge B false only in the former case (recall the reasoning underlining the Rewind History and Assume Independence attitudes).

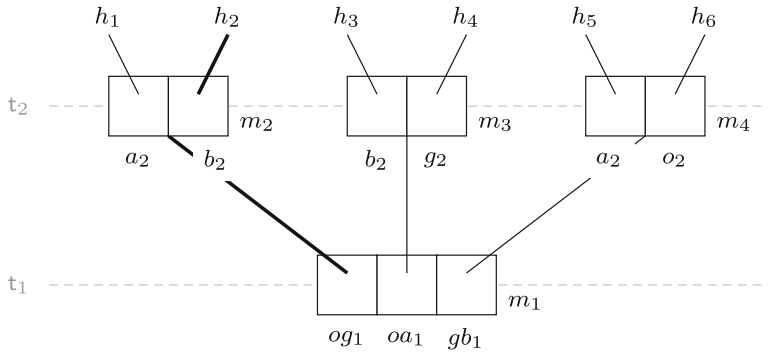

Fig. 5 Fruit basket example

Proposition 4 *The following principle is valid in any rewind model, but not valid in some independence model.*

$$(\mathsf{Exp}_{\Box})\quad \Box\neg\varphi \to ((\varphi \mathbin{\Box\!\!\rightarrow} \Box\psi) \to \Box(\varphi \mathbin{\Box\!\!\rightarrow} \psi))$$

Proof See Appendix B. □

Using item 2 in Corollary 2 and $\mathsf{Exp}_{\Box}$ we can show that $(\varphi \mathbin{\Box\!\!\rightarrow} \Box\psi) \to \Box(\varphi \mathbin{\Box\!\!\rightarrow} \psi)$ is valid in the class of rewind models. The validity of this principle can be proved directly from Definition 9, which ensures that the most similar φ-histories[25] to histories passing through a moment pass through the same moments. Note that the converse implication is not valid: suppose that we scheduled a lecture on Tuesday at 1pm and our default choice behavior is to follow the schedule. Then, "If I were not sick, I would be teaching" is settled true on Tuesday at 1pm, even though "If I were not sick, it would be settled that I would be teaching" may be false (*e.g.*, because there is a possibility that our bike breaks down on the way to school).

To see why the addition of the criterion regarding the number of independent events leads to the invalidity of $\mathsf{Exp}_{\Box}$, consider another example.

Example 3 Suppose that there is a basket containing an apple, a banana, an orange, and a grapefruit on a table. Next to the basket there is a jar containing three pieces of paper with the choices *orange+grapefruit*, *orange+apple*, *grapefruit+banana* written on them. Bob can pick one piece of paper and is given the fruits written on it. After Bob makes his choice, Ann can pick one of the remaining fruits from the basket. Assume that Bob picks the *orange+grapefruit*-paper and Ann picks the banana.

Example 3 is illustrated in Fig. 5. In the figure, Bob is agent 1 and his non-vacuous choices are og_1 (pick the *orange+grapefruit*-paper), oa_1 (pick the *orange+apple*-paper), and gb_1 (pick the *grapefruit+banana*-paper). Ann is agent 2 and her

[25] By "φ-history" we mean a history on which φ is true at the time of evaluation.

non-vacuous choices are a_2 (pick the apple), b_2 (pick the banana), g_2 (pick the grapefruit), and o_2 (pick the orange). The actual history (thick line) is h_2. In our terminology, both Bob and Ann are unconstrained agents—none of their actions are deviant. At m_2, there are no citrus fruits in the basket. But what if there were? According to Definition 10, the most similar history to h_2 satisfying this condition is h_3, where Bob picks the *orange+apple*-paper and Ann picks the banana—as she does at m_2/h_2. At t_2/h_3 it is settled that Ann can pick a banana, so "If there was a citrus fruit in the basket, it would be settled that Ann could pick a banana" is true at m_2/h_2. But consider the index m_2/h_1 where Ann picks the apple instead of the banana. Again, what if there was a citrus fruit in the basket? Reasoning as before, the most similar history to h_1 satisfying this condition is h_5, where Bob picks the *grapefruit+banana*-paper and Ann picks the apple. Since there is no banana in the basket at t_2/h_5, "If there was a citrus fruit in the basket, Ann could pick a banana" is false at m_2/h_1, and so "It is settled that, if there was a citrus fruit in the basket, Ann could pick a banana" is false at m_2/h_2.

To conclude this section, let us highlight a potential problem for our proposal emerging from Fig. 5. We have seen that, according to Definition 10, h_3 is the most similar history to h_2 on which Bob does not choose the *orange+grapefruit*-paper. So, "If Bob had picked a different piece of paper, then Ann would pick the banana" is true at m_2/h_2. But this is a counterintuitive conclusion: if Bob had picked a different piece of paper, he might have picked the *grapefruit+banana*-paper, in which case Ann could not even pick a banana!

We view this as a modeling issue: since choosing a banana over an apple is not the same type of choice as choosing a banana over a grapefruit, the two choices should not be labeled the same way (see the discussion of menu dependence in rational choice theory [21, 28, 40]). If we change the labeling, then the weaker (and unproblematic) "If Bob had picked a different piece of paper, then Ann might pick the banana" is true at m_2/h_2.[26]

This suggests the introduction of the next condition: for all $i \in Ag$ and $m, m' \in Mom$,

1. *Identity of Overlapping Menus*: if $Acts_i^m \cap Acts_i^{m'} \neq \varnothing$, then $Acts_i^m = Acts_i^{m'}$.

According to this condition, if an agent has the same type of choice available at two different moments, then the menus of alternative choices available to the agent at

[26] To be sure, suppose that we label Ann's choice at t_2/h_3 as b_2' (choosing a banana over a grapefruit) instead of b_2 (choosing a banana over an apple). In addition, for simplicity, assume that every agent i has a vacuous choice vc_i at all moments after t_2. Then, it is not difficult to see that histories h_3, h_4, h_5, and h_6 are equally similar to h_2: these histories have the same past overlap with h_2 (they all branch off from h_2 at m_1); the same number of agents make them branch off from h_2 (namely 1, *i.e.*, Bob); the same number of independent events occur on them (namely the events corresponding to the agents' vacuous choices); finally, the same number of deviant actions are performed on them (namely 0). Since these are all the histories on which Bob picks a different piece of paper at t_1 and Ann picks a banana only on h_3, we indeed conclude that, if Bob had picked a different piece of paper, then Ann *might* have picked a banana—the unwanted conclusion that Ann *would* have picked a banana does not follow. (Of course, according to this reasoning, we should also replace the label a_2 at t_2/h_5 with a_2').

those moments must be the same. The model in Fig. 5 does not satisfy this condition because Ann has two different but overlapping menus at m_2 and m_3, that is, $\{a_2, b_2\}$ and $\{b_2, g_2\}$ respectively. Interestingly, as proved in Appendix B, $\mathsf{Exp}_{\Box}$ remains invalid in the class of independence models satisfying the condition of identity of overlapping menus. In fact, the countermodel presented there satisfies a stronger condition: for all $m, m' \in Mom$,

2. *Uniformity of menus*: if $\mathsf{t}_m = \mathsf{t}_{m'}$, then $Acts^m = Acts^{m'}$.

While the condition of identity of overlapping menus is a desirable condition, the condition of uniformity of menus is not: as illustrated by Example 3, depending on what happens at a moment, different actions may become executable in the future.

4 A Refinement: From Independence to Influence

The definitions of similarity we introduced in the previous Section differ in how they treat choices of unconstrained agents. Definition 10 can be understood as fixing the choices of unconstrained agents when reasoning about counterfactual situations. On the other hand, Definition 9 does not keep track of the actions of unconstrained agents on the actual history. Despite this difference, a crucial assumption that both definitions of similarity rely on is that the evaluation of choice-driven counterfactuals depends on the default choice behavior of the agents. Do these definitions still make sense when evaluating a choice-driven counterfactual on a history where one or more agents behaved deviantly in the past? Should we ignore any past deviation from default choice behavior or take it into account when evaluating a choice-driven counterfactual? Consider the following variant of our running example.

Example 4 Everything is as in Example 1, except that, besides the two biased coins, Max can also choose a fair coin—and he knows this. Max's choice rule is the same: choose the coin that makes David lose. After Max makes his choice (and flips his chosen coin), David can choose to either leave or stay and play another round of the game with Max. Suppose that David nominates Max and bets heads but Max makes a mistake and flips the fair coin, which, lucky for David, lands heads. Then David chooses to leave the game.

How should we evaluate the following counterfactual after David leaves the game?

($C2$) If David were to bet heads again, he would lose $(\mathsf{X}do(bh_1) \Box\!\!\rightarrow \mathsf{XX}L)$.

Figure 6 depicts the relevant part of a model representing Example 4 needed to evaluate conditional $C2$. Max's choice of flipping the fair coin is represented by the action type fc_2; David's choices to leave and to stay are represented by the action types l_1 and s_1 respectively. The fair coin from the first game lands heads on all the depicted histories h_1-h_9, while the fair coin from the second game lands heads on histories h_3 and h_7. The actual history is h_1 (the thick line), where Max mistakenly flips the fair coin (which lands heads) and David decides not to play another round of the game.

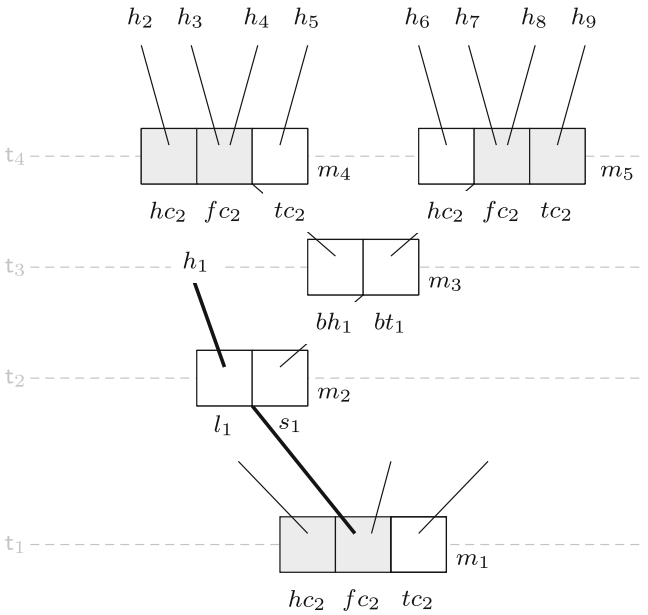

Fig. 6 The diagram represents a relevant portion of Example 4. After Max flips the fair coin at the end of the first game (moment m_1), David has a choice to either leave (l_1) or stay (s_1), followed by another round of the game described in Example 4

According to either Definition 9 or Definition 10, $C2$ is true at t_2/h_1: the most similar history to h_1 on which David bets heads during the second game is h_5, where $\mathsf{X}\mathsf{X}L$ is true at t_2.[27] It is not clear that this is the correct judgement about $C2$ given that Max mistakenly flipped the fair coin in the first game. The main issue is that neither definition of similarity takes into account the fact that the counterfactual is evaluated at a history along which Max acted deviantly. This raises a question about what Max would do in the second game. There are different ways to answer this question:

1. Forget that Max's actual choice was deviant and assume that he is still constrained by his choice rule (*i.e.*, he would flip the coin that makes David lose).
2. Assume that Max would make the same mistake and flip the fair coin.
3. Assume that Max would make *a* mistake, but we cannot tell which one (*e.g.*, he *might* flip the fair coin or the tails coin).
4. Assume that Max is no longer a constrained agent, so the only conclusion we can draw is that Max might flip *any* of the available coins.

[27] Note that $\mathsf{X}do(bh_1)$ is false at t_2/h_1. Of course, the successor of m_2 on h_1 is not represented in Fig. 6. It is assumed that the game is over at m_2/h_1 and so the next choice for David on h_1 does not involve betting heads.

Without further details about why Max made the deviant choice in the first game, it is not clear which of the above options is best. Perhaps Max made a fleeting mistake and there is no further explanation, which would suggest that option 1 is the best. There might be a systematic problem with the coins (*e.g.*, they are labeled incorrectly), which would suggest that either option 2 or option 3 is the best. Finally, options 4 is best if Max's deviant action is some type of signal that he is no longer being guided by his choice rule.

Remark 2 Counterfactuals like $C2$ play an important role in the analysis of strategic reasoning in game theory [7, 10, 39, 41, 43, 45, 54]. A central question in this literature is: What do the players expect that their opponents will do if an unexpected point in the game tree is reached? One answer (forward induction) is that players rationalize past behavior and use it as a basis for forming beliefs about future moves [3, 4, 47]. A second answer (backward induction) is that players ignore past behavior and reason only about their opponents' future moves [1, 9, 37, 47]. These different answers roughly correspond to the four different options listed above explaining Max's deviant choice. Forgetting that Max made a deviant choice and assuming he will be guided by his choice rule (option 1) is analogous to the assumptions underlying backward induction reasoning (the second answer). The other options can be viewed as different ways to rationalize Max's surprising choice, as in forward induction reasoning (the first answer).

In our framework, option 1 is implicitly assumed in both Definition 9 and Definition 10. Option 4 is best understood as Max transitioning from a constrained to an unconstrained agent, which requires a revision of Max's **dev** function. We leave the revision of the **dev** function to future work and suggest a way to represent options 2 and 3.

The reasoning underlying options 2 and 3 can be captured by generalizing Definition 10: When we suppose that David will bet tails, we follow the actual course of events up to the moment when David leaves the game, intervene on his choice by making sure that he will bet tails in the second game, fix all the actions of the unconstrained agents *and the fact that Max acted deviantly in the game*, and then let the future unfold according to the agents' default choice behavior. The key idea is that Max's deviant choice in the first game overrides his default behavior in the second game by fixing the fact that his choice will be deviant. Similarly, according to Definition 10, the choices of unconstrained agents are held fixed in counterfactual situations.

Both ideas can be captured by adding a relation between agent-moment pairs, where (i, m) is related to (j, m') means that i's choice at m *influences* j's choice at m': On the one hand, Max's deviant choice at m_1 influences him to make a deviant choice at m_4. On the other hand, Definition 10 requires that an unconstrained agent's choice at a moment m on a history h influences that agent to make the same type of choice at t_m on the most similar histories to h. This leads us to the following definitions.

Definition 11 (Influence relation) Let $\mathcal{F} = \langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ be an SLD_n frame and Ag be defined as above. An *influence relation* for Ag in $\mathcal{F}$ is a relation

$$\rightsquigarrow \subseteq (Ag \times Mom) \times (Ag \times Mom)$$

such that, for all $(i, m), (j, m') \in Ag \times Mom$,

1. $(i, m) \rightsquigarrow (j, m')$ whenever i and j are the same agent, who is unconstrained at both m and m', where m and m' occur at the same instant (*i.e.*, $i = j$, $i \in \underline{Ag}(m)$, $j \in \underline{Ag}(m')$, and $\mathsf{t}_m = \mathsf{t}_{m'}$);
2. otherwise, $(i, m) \rightsquigarrow (j, m')$ only if i and j are constrained agents at, respectively, m and m', where m' occurs either at the same time or later than m (*i.e.*, $i \notin \underline{Ag}(m)$, $j \notin \underline{Ag}(m')$, and $m'' \leq m'$ for some $m'' \in \mathsf{t}_m$).[28]

For $(i, m), (j, m') \in Ag \times Mom$, we will write $(i, m) \rightsquigarrow^I (j, m')$ when 1 holds and $(i, m) \rightsquigarrow^D (j, m')$ when 2 holds.

Definition 12 (Fixed actions) Let $\mathcal{F} = \langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ be an SLD_n frame and Ag be defined as above. For any index $m/h \in Ind$ and agent $i \in Ag$ let

$$\mathbf{fixed}_i^{m/h} : Ag \times Mom \to 2^{Acts}$$

be defined as follows:

1. $\mathbf{fixed}_i^{m/h}(j, m') = \{\mathbf{act}(m/h)(i)\}$ if $(i, m) \rightsquigarrow^I (j, m')$ and $\mathbf{act}(m/h)(i) \in Acts_j^{m'}$;
2. $\mathbf{fixed}_i^{m/h}(j, m') \subseteq Acts_j^{m'} \cap \mathbf{dev}(m')$ if $(i, m) \rightsquigarrow^D (j, m')$ and $\mathbf{act}(m/h)(i) \in \mathbf{dev}(m)$;
3. $\mathbf{fixed}_i^{m/h}(j, m') = \varnothing$ otherwise.

When $(i, m) \rightsquigarrow (j, m')$, then $\mathbf{fixed}_i^{m/h}(j, m')$ are the actions that are "fixed" for j at m' given the influence of (i, m) on (j, m') and the action that i performs at m/h. In case 1 (where $i = j$ is unconstrained at both m and m' occurring at the same time), the fixed action is the action type that is performed by agent i at m/h. In case 2 (where i and j are constrained agents), what is fixed is the fact that agent j will choose a deviant action at m', given that i's action at m/h was deviant.

We can now refine the function *n_indep* in a natural way:

$$n\_indep^*(h, h') = \sum_{m \in h} |\{(j, m') \in Ag \times h' \mid \text{there is } i \in Ag \text{ s.t. } (i, m) \rightsquigarrow (j, m') \text{ and } \mathbf{act}(m'/h')(j) \in \mathbf{fixed}_i^{m/h}(j, m')\}|$$

The function $n\_indep^*(h, h')$ counts the number of agents on h' that are "properly influenced" by agents on h, in the sense that they perform an action from their

[28] One can imagine generalizations of this definition where, for instance, unconstrained agents may influence constrained agents. However, we use this simpler definition since it covers the cases that we have in mind for this paper; a full analysis of influence would require a separate paper.

set of fixed actions (if non-empty). It is not difficult to see that, if $\rightsquigarrow^D = \varnothing$, then $n\_indep^*(h, h') = n\_indep(h, h')$.

Going back to Example 4, suppose that $(2, m_1) \rightsquigarrow (2, m_4)$ and $\mathbf{fixed}_2^{m_1/h_1}(2, m_4) = Acts_2^{m_4} \cap \mathbf{dev}(m_4)$ (in line with option 3 above). That is, if 2 chooses deviantly at m_1, then 2 will choose deviantly at m_4. Then, $n\_indep^*(h_1, h_5) < n\_indep^*(h_1, h_2) = n\_indep^*(h_1, h_3) = n\_indep^*(h_1, h_4)$, since 2 chooses deviantly at m_4 on all of h_2, h_3, and h_4 (as 2 does at m_1 on h_1) but not at m_4 on h_5. Hence, histories h_2, h_3, and h_4 are more similar to h_1 than h_5, and so the counterfactual $C2$ is false at m_2/h_1 according to Definition 10 using n_indep^* in place of n_indep.

5 Conclusion

In this paper, we studied the semantics and logical properties of choice-driven counterfactuals in a stit logic with action types, instants and deviant choices. Following Lewis [30], we interpreted counterfactual statements using a relation of relative similarity on histories. We introduced two definitions of similarity motivated by different intuitions about how choice rules guide the agents' actions in counterfactual situations: the Rewind History intuition and the Assume Independence intuition. We showed how to adapt our definitions to situations in which some agents perform a deviant action. We have highlighted the subtle issues that arise when merging a logic of counterfactuals with a logic of branching time and agency.

There are a number of interesting technical questions that arise concerning our full language $\mathcal{L}_{\mathsf{SLD}_n}^{\Box\!\!\rightarrow}$. One question concerns whether $\mathcal{L}_{\mathsf{SLD}_n}^{\Box\!\!\rightarrow}$ is strictly more expressive than $\mathcal{L}_{\mathsf{SLD}_n}$ over our class of models. For instance, consider the formula $\neg\varphi \Box\!\!\rightarrow \bot$, which says that φ is true at all indices occurring at the instant of evaluation (cf. [30, p. 22]). Note that at any index m/h in any model $\mathcal{M}$ there is an $n \in \mathbb{N}$ such that $m \in succ^n(m_0)$. This means that $\mathcal{M}, m/h \models \neg\varphi \Box\!\!\rightarrow \bot$ iff $\mathcal{M}, m/h \models \mathsf{Y}^n\Box\mathsf{X}^n\varphi$. Thus, in any model and index we can find a formula of $\mathcal{L}_{\mathsf{SLD}_n}$ that is equivalent to $\neg\varphi \Box\!\!\rightarrow \bot$ at that index. Of course, n (and, hence, the formula of $\mathcal{L}_{\mathsf{SLD}_n}$) varies depending on the index. This suggests that comparing the expressive power of $\mathcal{L}_{\mathsf{SLD}_n}^{\Box\!\!\rightarrow}$ and $\mathcal{L}_{\mathsf{SLD}_n}$ over our models is not straightforward.

A second question concerns the possibility of a sound and complete axiomatization of rewind (resp. independence) models with respect to our full language. We do have a sound and complete axiomatization of SLD_n frames (Definition 4) in a language without counterfactuals (Theorem 1). For our full language, we identified some core validities (Proposition 2 and Proposition 3) and an interesting formula that distinguishes rewind and independence models (Proposition 4). Since our definitions of similarity (Definition 9 and Definition 10) involve counting (deviant) actions along different histories, we expect that a complete axiomatization (if there is one) will require an extension of our language.

Another direction for future research is to explore applications of the logical framework developed in this paper. Branching-time logics with both agency operators and counterfactuals are a powerful tool to reason about complex social interactions.

In particular, logics of this sort seem to be necessary to clarify complex moral and legal ideas, such as the concept of responsibility [2, 11, 12, 20, 32] and "could have done otherwise" [5]. In addition, the discussion in Section 4 and Remark 2 suggests that a stit logic with counterfactuals may be fruitfully used to incorporate strategic reasoning in stit, thus advancing recent research connecting stit and game-theory, see, *e.g.*, [19, 29, 48, 51]). We conjecture that the latter application may call for a framework combining our approach to the semantics of counterfactuals with extensions of stit logics with epistemic operators [23, 27, 50] and probabilistic belief operators [13].

Appendix: A Completeness of SLD_n

In this appendix we prove that the axiom system SLD_n is complete with respect to the class of all SLD_n frames.[29] The proof consists of two parts. First, we show that SLD_n is sound and complete with respect to a class of Kripke models (called *pseudo-models*). By elaborating on a technique presented by [24], we then prove that every pseudo-model in which a formula $\varphi \in \mathcal{L}_{\mathsf{SLD}}$ is satisfiable can be turned into an SLD_n model in which φ is satisfiable.

A.1 Pseudo-Models

Pseudo-models consist of a non-empty set W of possible states representing moment-history pairs partitioned into equivalence classes by an equivalence relation $R_\Box$. Intuitively, every equivalence class of $R_\Box$ represents a *moment*. Besides $R_\Box$, pseudo-models feature the following elements: two accessibility relations, denoted R_X and R_Y, modeling, respectively, what happens next and what happened a moment ago; a function f_{do} assigning to every possible state the profile that is performed at that state; finally, a function f_{dev} assigning to every state a set of deviant individual actions.

Remark 3 We adopt the following standard notation. For any set S, element $s \in S$, and relation $R \subseteq S \times S$, $R(s) = \{s' \in S \mid sRs'\}$. For any number $n \in \mathbb{N}$, $R^n \subseteq S \times S$ is defined recursively by setting: wR^0v iff $w = v$; $wR^{n+1}v$ iff there is $u \in S$ s.t. wR^nu and uRv.

Definition 13 (Pseudo-model) A pseudo-model is a tuple $\langle W, R_\Box, R_\mathsf{X}, R_\mathsf{Y}, f_{do}, f_{dev}, v\rangle$, where $W \neq \varnothing$, $R_\Box$ is an equivalence relation on W, R_X and R_Y are binary relations on W, $f_{do} : W \to Ag\text{-}Acts$ is the *action function*, $f_{dev} : W \to 2^{Acts}$ is the *deviant-choice function*, and $v : \mathrm{Prop} \to 2^W$ is a valuation function. For any $w \in W$ and $i \in \mathrm{Ag}$, let:

[29] The proof that SLD_n is sound with respect to the class of all SLD_n frames is a matter of routine validity check and it is thus omitted.

$Acts_i^w = \bigcup\{f_{do}(w')(i) \in Acts_i \mid w' \in R_\Box(w)\}$ be the actions *available to agent i at* $R_\Box(w)$;
$Acts^w = \bigcup_{i \in Ag} Acts_i^w$ be the individual actions *executable at* $R_\Box(w)$.

Define $R_{Ag} \subseteq W \times W$ by setting: for all $w, w' \in W$, $wR_{Ag}w'$ iff $wR_\Box w'$and $f_{do}(w) = f_{do}(w')$. The elements of a pseudo-model are assumed to satisfy the following conditions:

1. Properties of R_X and R_Y: for all $w, w_1, w_2 \in W$,
   - 1.1. *Seriality of* R_X: there is $w' \in W$ such that $wR_\mathsf{X}w'$.
   - 1.2. R_X*-functionality*: if $wR_\mathsf{X}w_1$ and $wR_\mathsf{X}w_2$, then $w_1 = w_2$.
   - 1.3. R_Y*-functionality*: if $wR_\mathsf{Y}w_1$ and $wR_\mathsf{Y}w_2$, then $w_1 = w_2$.
   - 1.4. *Converse*: $w_1R_\mathsf{Y}w_2$ iff $w_2R_\mathsf{X}w_1$.
2. *Independence of Agents*: for all $w \in W$ and $\alpha \in Ag\text{-}Acts$, if $\alpha(j) \in Acts^w$ for all $j \in Ag$, then there is $w' \in R_\Box(w)$ such that $f_{do}(w') = \alpha$.
3. *No Choice between Undivided Histories*: for all $w_1, w_2, w_3 \in W$, if $w_1R_\mathsf{X}w_2$ and $w_2R_\Box w_3$, then there is $v \in W$ such that $w_1R_{\mathrm{Ag}}v$ and $vR_\mathsf{X}w_3$.
4. *Properties of* f_{dev}: for all $w, w' \in W$ and $i \in Ag$,
   - 4.1. *Moment-invariance*: if $wR_\Box w'$, then $f_{dev}(w) = f_{dev}(w')$.
   - 4.2. *Executability of Deviant Actions*: $f_{dev}(w) \subseteq Acts^w$.
   - 4.3. *Availability of Non-deviant Actions*: $Acts_i^w \setminus f_{dev}(w) \neq \varnothing$.
   - 4.4. *(In)determinism of Choice Rules*: if $Acts_i^w \cap f_{dev}(w) \neq \varnothing$, then $|Acts_i^w \setminus f_{dev}(w)| = 1$.

Definition 14 (Truth for $\mathcal{L}_{\mathsf{SLD}_n}$ in a pseudo-model) Given a pseudo-model M, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{SLD}_n}$ at a state w in M, denoted $M, w \models \varphi$, is defined recursively. Truth of atomic propositions and the Boolean connectives is defined as usual. The remaining clauses are as follows: where $\blacksquare \in \{\Box, \mathsf{X}, \mathsf{Y}\}$,

$$
\begin{aligned}
M, w &\models do(a_i) \quad \text{iff } f_{do}(w)(i) = a_i \\
M, w &\models dev(a_i) \quad \text{iff } a_i \in f_{dev}(w) \\
M, w &\models \blacksquare\varphi \quad \text{iff for all } w' \in W, \text{if } wR_\blacksquare w', \text{then } M, w' \models \varphi
\end{aligned}
$$

Theorem 2 *The axiom system* SLD_n*, defined by the axioms and rules in* Table 2*, is sound and complete with respect to the class of all pseudo-models.*

The proof of Theorem 2 is entirely standard: soundness is proved via a routine validity check and completeness is proved via the construction of a canonical model for SLD_n (see [8, Chapter 4.2]). We only provide the definition of the canonical model for SLD_n and leave the rest to the reader. Let $\mathcal{W}$ be the set of all maximal consistent sets of SLD_n. Where $w \in \mathcal{W}$ and $\blacksquare \in \{\Box, \mathsf{X}, \mathsf{Y}\}$, define $w/\blacksquare = \{\varphi \in \mathcal{L}_{\mathsf{SLD}_n} \mid \blacksquare\varphi \in w\}$.

Definition 15 The canonical SLD_n model is a tuple $\langle W^c, R^c_\Box, R^c_\mathsf{X}, R^c_\mathsf{Y}, f^c_{do}, f^c_{dev}, v^c\rangle$, where

- $W^c = \mathcal{W}$ and $v^c : Prop \to 2^{W^c}$ is s.t., for all $w \in W^c$, $w \in v^c(p)$ iff $p \in w$;
- where $\blacksquare \in \{\Box, \mathsf{X}, \mathsf{Y}\}$, $R^c_\blacksquare \subseteq W^c \times W^c$ is s.t., for all $w, w' \in W^c$, $wR^c_\blacksquare w'$ iff $w/\blacksquare \subseteq w'$;
- $f^c_{do} : W^c \to Ag\text{-}Acts$ is s.t., for all $w \in W^c$, $f^c_{do}(w) = \alpha$ iff $do(\alpha) \in w$;
- $f^c_{dev} : W^c \to 2^{Acts}$ is s.t., for all $w \in W^c$ and $a_i \in Acts$, $a_i \in f^c_{dev}(w)$ iff $dev(a_i) \in w$.

A.2 From Pseudo-Models to SLD_n Models

Call a *pointed pseudo-model* any pair M, w such that M is a pseudo-model and w a state in M. By Theorem 2, for any SLD_n-consistent formula φ, there is a pointed pseudo-model M, w such that $M, w \models \varphi$. We want to show that M can be transformed into an SLD_n model in which φ is satisfiable. To build stit models from Kripke models similar to our pseudo-models, Herzig and Lorini [24] use a construction consisting of two preliminary steps: (1) the relevant Kripke model is *unraveled*[30] in order to ensure that the relation R_X generates a treelike ordering of the equivalence classes of $R_\Box$ (recall that these represent moments); (2) from a certain point on along the relation R_X in the unraveled model, every equivalence class of $R_\Box$ is forced to be a singleton. Step (2) guarantees that there is a one-to-one correspondence between states in the unraveled model and indices in the stit model built from it. The presence of the operator Y in the language of SLD_n requires us to refine the unraveling procedure in step (1). We present the said refinement in details (Steps 1 and 2 below) and only sketch the rest of the proof (Steps 3 to 4 below), which proceeds (except for a few minor modifications) as in [18, Appendix A.1.2].

Step 1: Extended language and complexity measures

Our first task is to define an unraveling procedure $\mathbf{u}$ that takes a pointed pseudo-model M, w and a formula $\varphi \in \mathcal{L}_{\mathsf{SLD}_n}$ and returns a pointed pseudo-model $\mathbf{u}^\varphi(M, w)$ satisfying:

$(P1)$ $\quad M, w \models \varphi$ iff $\mathbf{u}^\varphi(M, w) \models \varphi$.

The idea is roughly as follows: we first identify the earliest state w' needed to determine whether φ is true at w; then, we unravel R_X around the $R_\Box$-equivalence class of w'. To make this work, we need to extend our language and introduce three complexity measures of the formulas in the extended set $\mathcal{L}'_{\mathsf{ALD}}$: (i) the Y-*depth of* φ is needed to identify w' and the state corresponding to w in the unraveled model;

[30] A standard definition of unraveling can be found in [8, p. 63]. Herzig and Lorini's [24] definition is a generalization of the latter definition.

(ii) the *size of* φ and (iii) the *c-size of* φ are needed to define a well-founded strict partial order $<_c^S$ on $\mathcal{L}'_{\mathsf{ALD}}$. The proof that our unraveling procedure satisfies $P1$ will be on $<_c^S$-induction on φ (cf. Proposition 6).

Definition 16 (Extended language) Let $Prop$ and $Acts$ be as before. The set $\mathcal{L}'_{\mathsf{SLD}_n}$ is generated by the following grammar:

$$p \mid do(a_i) \mid dev(a_i) \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid \mathsf{X}\varphi \mid \mathsf{Y}\varphi \mid \boxplus\varphi$$

where $p \in Prop$ and $a_i \in Acts$.

The evaluation rule for $\boxplus\varphi$ in the class of pseudo-models is as follows:

$$M, w \models \boxplus\varphi \text{ iff for all } v, u \in W \text{ s.t. } wR_{\mathsf{X}}v \text{ and } vR_{\mathsf{Y}}u,\ M, u \models \varphi$$

Accordingly, $\boxplus\varphi \leftrightarrow \mathsf{XY}\varphi$ and $\boxplus\varphi \leftrightarrow \varphi$ are valid on all pseudo-models.

Definition 17 (Y-*depth* of $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$) The Y-depth $d(\varphi)$ of $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$ is defined as:

$$\begin{aligned} d(p) = d(do(a_i)) = d(dev(a_i)) &= 0 \\ d(\neg\varphi) = d(\Box\varphi) = d(\mathsf{X}\varphi) &= d(\varphi) \\ d(\boxplus\varphi) = d(\mathsf{Y}\varphi) &= d(\varphi) + 1 \\ d(\varphi \wedge \psi) &= max\{d(\varphi), d(\psi)\} \end{aligned}$$

Definition 18 (*Size* of $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$) The size $S(\varphi)$ of $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$ is defined as:

$$\begin{aligned} S(p) = S(do(a_i)) = S(dev(a_i)) &= 1 \\ S(\neg\varphi) = S(\boxplus\varphi) &= S(\varphi) + 1 \\ S(\Box\varphi) = S(\mathsf{X}\varphi) = S(\mathsf{Y}\varphi) &= S(\varphi) + 2 \\ S(\varphi \wedge \psi) &= S(\varphi) + S(\psi) + 1 \end{aligned}$$

Definition 19 (*c-size* of $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$) The c-size $c(\varphi)$ of $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$ is defined as:

$$\begin{aligned} c(p) = c(do(a_i)) = c(dev(a_i)) &= 0 \\ c(\neg\varphi) = c(\Box\varphi) = c(\mathsf{X}\varphi) = c(\mathsf{Y}\varphi) = c(\boxplus\varphi) &= c(\varphi) \\ c(\varphi \wedge \psi) &= c(\varphi) + c(\psi) + 1 \end{aligned}$$

Definition 20 For any $\varphi, \psi \in \mathcal{L}'_{\mathsf{SLD}_n}$, we set: $\varphi <_c^S \psi$ iff either $c(\varphi) < c(\psi)$ or $(c(\varphi) = c(\psi)$ and $S(\varphi) < S(\psi))$.

Lemma 1 *$<_c^S$ is a well-founded strict partial order between the formulas of $\mathcal{L}'_{\mathsf{SLD}_n}$.*

Proof Straightforward from Def. 20. □

Lemma 2 *For any $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$ and $n \in \mathbb{N}$ such that $n \geq d(\varphi)$, there is $\varphi' \in \mathcal{L}'_{\mathsf{SLD}_n}$ s.t. (1) $\varphi \leftrightarrow \varphi'$ is valid on any pseudo-model, (2) $d(\varphi') = n$, and (3) $c(\varphi') = c(\varphi)$.*

Proof Immediate from the fact that $\varphi \leftrightarrow \boxplus\varphi$ is valid on any pseudo-model, that $d(\boxplus\varphi) = d(\varphi) + 1$, and that $c(\boxplus\varphi) = c(\varphi)$. □

Step 2: Unraveling procedure

We adopt the following notation: where M, w is a pointed pseudo-model and $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$,

1. $d(w, \varphi)$ is the greatest number n satisfying: $n \leq d(\varphi)$ and there is a $v \in W$ such that $wR^n_{\mathsf{Y}}v$ (equiv. $vR^n_{\mathsf{X}}w$);
2. where $n = d(w, \varphi)$, $s_{(w,\varphi)} \in W$ is the state v satisfying: $wR^n_{\mathsf{Y}}v$ (equiv. $vR^n_{\mathsf{X}}w$).[31]

Definition 21 ($d(w, \varphi)$-unraveling) Let M, w be a pointed pseudo-model and $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$. The $d(w, \varphi)$-unraveling of M, w is the tuple

$$M^{(w,\varphi)} = \left\langle W^{(w,\varphi)}, R^{(w,\varphi)}_{\Box}, R^{(w,\varphi)}_{\mathsf{X}}, R^{(w,\varphi)}_{\mathsf{Y}}, f^{(w,\varphi)}_{do}, f^{(w,\varphi)}_{dev}, \nu^{(w,\varphi)} \right\rangle$$

where:

- $W^{(w,\varphi)}$ is the set of all sequences $\overrightarrow{w_n} = w_1 w_2 \dots w_n$ s.t. $w_1 R_{\Box} s_{(w,\varphi)}$ and $w_i R_{\mathsf{X}} w_{i+1}$, where $1 \leq i < n$;
- $R^{(w,\varphi)}_{\Box} \subseteq W^{(w,\varphi)} \times W^{(w,\varphi)}$ is s.t. $\overrightarrow{w_n} R^{(w,\varphi)}_{\Box} \overrightarrow{v_m}$ iff $n = m$, $w_i R_{\Box} v_i$ for $i \leq n$, and $f_{do}(w_i) = f_{do}(v_i)$ for $i < n$;
- $R^{(w,\varphi)}_{\mathsf{X}} \subseteq W^{(w,\varphi)} \times W^{(w,\varphi)}$ is s.t. $\overrightarrow{w_n} R^{(w,\varphi)}_{\mathsf{X}} \overrightarrow{v_m}$ iff $\overrightarrow{v_m} = \overrightarrow{w_n} v_m$ and $w_n R_{\mathsf{X}} v_m$;
- $R^{(w,\varphi)}_{\mathsf{Y}} \subseteq W^{(w,\varphi)} \times W^{(w,\varphi)}$ is s.t. $\overrightarrow{w_n} R^{(w,\varphi)}_{\mathsf{Y}} \overrightarrow{v_m}$ iff $\overrightarrow{w_n} = \overrightarrow{v_m} w_n$ and $w_n R_{\mathsf{Y}} v_m$;
- $f^{(w,\varphi)}_{do} : W^{(w,\varphi)} \to Ag\text{-}Acts$ is s.t. $f^{(w,\varphi)}_{do}(\overrightarrow{w_n}) = f_{do}(w_n)$
- $f^{(w,\varphi)}_{dev} : W^{(w,\varphi)} \to 2^{Acts}$ is s.t. $f^{(w,\varphi)}_{dev}(\overrightarrow{w_n}) = f_{dev}(w_n)$
- $\nu^{(w,\varphi)} : Prop \to 2^{W^{(w,\varphi)}}$ is s.t. $\overrightarrow{w_n} \in \nu^{(w,\varphi)}(p)$ iff $w_n \in \nu(p)$.

Let $\sigma_{(w,\varphi)}$ be the sequence $w_1 w_2 \dots w_n$ s.t. $w_1 = s_{(w,\varphi)}$, $w_n = w$, and $n = d(w, \varphi) + 1$.[32]

Remark 4 The construction of $M^{(w,\varphi)}$ and $\sigma_{(w,\varphi)}$ ultimately depends on w and $d(\varphi)$. Hence, if $d(\varphi) = d(\psi)$, then $M^{(w,\varphi)} = M^{(w,\psi)}$ and $\sigma(w, \varphi) = \sigma(w, \psi)$.

We will use the following lemma repeatedly later on.

Lemma 3 *Let M, w be a pointed pseudo-model and $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$. For all $\overrightarrow{w_n}$ in $M^{(w,\varphi)}$ and v in M, if $w_n R_{\Box} v$, then there is $\overrightarrow{v_n}$ in $M^{(w,\varphi)}$ such that $v_n = v$ and $\overrightarrow{w_n} R^{(w,\varphi)}_{\Box} \overrightarrow{v_n}$.*

[31] Observe that the uniqueness of such state is guaranteed by the functionality of R_{Y}.

[32] The existence of $\sigma_{(w,\varphi)}$ is guaranteed by the way $W^{(w,\varphi)}$ is built and the uniqueness of $\sigma_{(w,\varphi)}$ by the functionality of R_{X}.

Proof The proof proceeds by an easy induction on n and relies on the fact that M satisfies condition 3 from Def. 13 (*i.e.*, no choice between undivided histories). □

Proposition 5 *For any pointed pseudo-model M, w and $\varphi \in \mathcal{L}'_{SLD_n}$, $M^{(w,\varphi)}$ is a pseudo-model.*

Proof The proof follows immediately from Def. 21 and from the fact that M is a pseudo-model. To illustrate, we check that $M^{(w,\varphi)}$ satisfies condition 2 from Def. 13 (*i.e.*, independence of agents): Let $\overrightarrow{w_n} \in W^{(w,\varphi)}$ and $\alpha \in$ Ag-Acts be s.t., for all $j \in$ Ag, there is $\overrightarrow{v_{n_j}} \in R_{\Box}^{(w,\varphi)}(\overrightarrow{w_n})$ s.t. $f_{do}^{(w,\varphi)}(\overrightarrow{v_{n_j}})(j) = \alpha(j)$. Then, by the def. of $R_{\Box}^{(w,\varphi)}$ and $f_{do}^{(w,\varphi)}$, for all $j \in$ Ag, there is v_{n_j} s.t. $v_{n_j} \in R_{\Box}(w_n)$ and $f_{do}(v_{n_j})(j) = \alpha(j)$. Since M satisfies condition 2 from Def. 13, it follows that there is $u \in R_{\Box}(w_n)$ s.t. $f_{do}(u) = \alpha$. By Lem. 3 and the def. of $f_{do}^{(w,\varphi)}$, we conclude that there is $\overrightarrow{u_n} \in W^{(w,\varphi)}$ s.t. $u_n = u$, $\overrightarrow{u_n} R_{\Box}^{(w,\varphi)} \overrightarrow{w_n}$, and $f_{do}^{(w,\varphi)}(\overrightarrow{u_n}) = f_{do}(u) = \alpha$. □

The next three lemmas will be key to prove Proposition 6 below.

Lemma 4 *Let M, w be a pointed pseudo-model. For any $v \in W$ and $\varphi, \psi \in \mathcal{L}'_{SLD_n}$, if $w R_{\Box} v$ and $d(\varphi) = d(\psi)$, then (1) $M^{(w,\varphi)} = M^{(v,\psi)}$ and (2) $\sigma_{(w,\varphi)} R_{\Box}^{(w,\varphi)} \sigma_{(v,\psi)}$.*

Proof (1) It is not difficult to see that, by condition 3 from Def. 13 (*i.e.*, no choice between undivided histories), if $w R_{\Box} v$ and $d(\varphi) = d(\psi)$, then $d(w, \varphi) = d(v, \psi)$ and $s_{(w,\varphi)} R_{\Box} s_{(v,\psi)}$. In tandem with Def. 21, the latter fact entails that $M^{(w,\varphi)} = M^{(v,\psi)}$. (2) Since the last element of $\sigma(w, \varphi)$ is w and $w R_{\Box} v$, it follows from Lem. 3 that there is $\overrightarrow{v_n} \in W^{(w,\varphi)}$ s.t. $v_n = v$ and $\sigma(w, \varphi) R_{\Box}^{(w,\varphi)} \overrightarrow{v_n}$ (so $n = d(w, \varphi) + 1 = d(v, \psi) + 1$). By the def. of $\sigma(v, \psi)$ and the functionality of R_{X}, this entails that $\overrightarrow{v_n} = \sigma(v, \psi)$. Hence, $\sigma(w, \varphi) R_{\Box}^{(w,\varphi)} \sigma(v, \psi)$. □

Lemma 5 *Let M, w be a pointed pseudo-model. For any $v \in W$ and $\varphi, \psi \in \mathcal{L}'_{SLD_n}$, if $w R_{\mathsf{X}} v$ and $d(\psi) = d(\varphi) + 1$, then (1) $M^{(w,\varphi)} = M^{(v,\psi)}$ and (2) $\sigma(w, \varphi) R_{\mathsf{X}}^{(w,\varphi)} \sigma(v, \psi)$.*

Proof (1) It is not difficult to see that, by the def. of $d(w, \varphi)$ and $s_{(w,\varphi)}$ and the functionality of R_{X} and R_{Y}, if $w R_{\mathsf{X}} v$ and $d(\psi) = d(\varphi)+1$, then $d(v, \psi) = d(w, \varphi)+1$ and $S_{(w,\varphi)} = S_{(v,\psi)}$. Given Def. 21, the latter fact entails that $M^{(w,\varphi)} = M^{(v,\psi)}$. (2) Immediate since $\sigma(w, \varphi)$ and $\sigma(v, \psi)$ have the same initial state, R_{X} is functional, and $w R_{\mathsf{X}} v$. □

Lemma 6 *Let M, w be a pointed pseudo-model. For any $\varphi, \psi \in \mathcal{L}'_{\mathsf{SLD}_n}$, if $d(\psi) = d(\varphi) + 1$, then $M^{(w,\psi)}, \sigma_{(w,\psi)} \models \varphi$ iff $M^{(w,\varphi)}, \sigma_{(w,\varphi)} \models \varphi$.*

Proof If $d(w, \psi) = d(w, \varphi)$, then $M^{(w,\psi)} = M^{(w,\varphi)}$ and $\sigma_{(w,\psi)} = \sigma_{(w,\varphi)}$ by Def. 21, whence the result. If $d(w, \psi) \neq d(w, \varphi)$, then $d(w, \psi) = d(w, \varphi) + 1$ by the def. of $d(w, \varphi)$. Let $n = d(w, \varphi)$, so that $d(w, \psi) = n + 1$. By the def. of $s_{(w,\varphi)}$

and $s_{(w,\psi)}$, $s_{(w,\varphi)}R_{\mathsf{X}}^{n}w$ and $s_{(w,\psi)}R_{\mathsf{X}}^{n+1}w$, and so $s_{(w,\psi)}R_{\mathsf{X}}s_{(w,\varphi)}$ by the functionality of R_{X}. Consider now $M^{(w,\psi)}$. It is easy to check that the two-element sequence $s_{(w,\psi)}s_{(w,\varphi)}$ is s.t. (1) $s_{(w,\psi)}s_{(w,\varphi)} \in W^{(w,\psi)}$ and (2) $S_{(w,\psi)}S_{(w,\varphi)}(R_{\mathsf{X}}^{(w,\varphi)})^{n}\sigma(w,\psi)$. Let

$$M^{(w,\psi,\varphi)} = \left\langle W^{(w,\psi,\varphi)}, R_{\Box}^{(w,\psi,\varphi)}, R_{\mathsf{X}}^{(w,\psi,\varphi)}, R_{\mathsf{Y}}^{(w,\psi,\varphi)}, f_{do}^{(w,\psi,\varphi)}, f_{dev}^{(w,\psi,\varphi)}, \nu^{(w,\psi,\varphi)}\right\rangle$$

be the submodel of $M^{(w,\psi)}$ generated by $s_{(w,\psi)}s_{(w,\varphi)}$ via $R_{\Box}^{(w,\psi)}$ and $R_{\mathsf{X}}^{(w,\psi)}$.[33] Obviously, $\sigma_{(w,\psi)} \in W^{(w,\psi,\varphi)}$. In addition, since $M^{(w,\psi,\varphi)}$ is obtained by "cutting" $M^{(w,\psi)}$ in the past taking into account the Y-depth of φ (recall that $n = d(w,\varphi)$), we have that:

1. $M^{(w,\psi)}, \sigma_{(w,\psi)} \models \varphi$ iff $M^{(w,\psi,\varphi)}, \sigma_{(w,\psi)} \models \varphi$.

Now, define a mapping $f : W^{(w,\psi,\varphi)} \to W^{(w,\varphi)}$ by setting: for every $\overrightarrow{w_m} \in W^{(w,\psi,\varphi)}$, $f(\overrightarrow{w_m}) = w_2w_3\ldots w_m$. That is, $f(\overrightarrow{w_m})$ is the sequence obtained by eliminating the first element of $\overrightarrow{w_m}$. We now prove the following facts:

2. for all $\overrightarrow{w_m} \in W^{(w,\psi,\varphi)}$, $f(\overrightarrow{w_m}) \in W^{(w,\varphi)}$;
3. $f(\sigma_{(w,\psi)}) = \sigma_{(w,\varphi)}$;
4. the function f is a bounded morphism from $M^{(w,\psi,\varphi)}$ to $M^{(w,\varphi)}$. □

Proof of 2. Let $\overrightarrow{w_m} = w_1w_2\ldots w_m \in W^{(w,\psi,\varphi)}$, so $f(\overrightarrow{w_m}) = w_2\ldots w_m$. By the def. of $W^{(w,\psi,\varphi)}$, for all i s.t. $2 \leq i < m$, $w_iR_{\mathsf{X}}w_{i+1}$. In addition, since $M^{(w\psi,\varphi)}$ is generated by $S_{(w,\psi)}S_{(w,\varphi)}$ via $S_{(w,\psi)}$ and $S_{(w,\psi)}$, $w_1w_2S_{(w,\psi)}S_{(w,\psi)}S_{(w,\varphi)}$. Hence, $w_2R_{\Box}S_{(w,\varphi)}$. □

Proof of 3. Straightforward from the def. of $\sigma_{(w,\varphi)}$ and $\sigma_{(w,\psi)}$, since $s_{(w,\psi)}R_{\mathsf{X}}s_{(w,\varphi)}$. □

Proof of 4. Let $\blacksquare \in \{\Box, \mathsf{X}, \mathsf{Y}\}$. We need to prove that, for all $\overrightarrow{w_n}, \overrightarrow{v_m} \in W^{(w,\psi,\varphi)}$, $\overrightarrow{u_k} \in W^{(w,\varphi)}$, $a_i \in Acts$, and $p \in Prop$,

4.1 if $\overrightarrow{w_n}R_{\blacksquare}^{(w,\psi,\varphi)}\overrightarrow{v_m}$, then $f(\overrightarrow{w_n})R_{\blacksquare}^{(w,\varphi)}f(\overrightarrow{v_m})$;
4.2 if $f(\overrightarrow{w_n})R_{\blacksquare}^{(w,\varphi)}\overrightarrow{u_k}$, then there is $\overrightarrow{v_m} \in W^{(w,\psi,\varphi)}$ s.t. $f(\overrightarrow{v_m}) = \overrightarrow{u_k}$ and $\overrightarrow{w_n}R_{\blacksquare}^{(w,\psi,\varphi)}\overrightarrow{v_m}$;
4.3 $f_{do}^{(w,\psi,\varphi)}(\overrightarrow{w_n}) = f_{do}^{(w,\varphi)}(f(\overrightarrow{w_n}))$ and $f_{dev}^{(w,\psi,\varphi)}(\overrightarrow{w_n}) = f_{dev}^{(w,\varphi)}(f(\overrightarrow{w_n}))$;
4.5 $\overrightarrow{w_n} \in \nu^{(w,\psi,\varphi)}(p)$ iff $f(\overrightarrow{w_n}) \in \nu^{(w,\varphi)}(p)$.

[33] That is, $M^{(w,\psi,\varphi)}$ is the smallest submodel of $M^{(w,\varphi)}$ such that (1) $S_{(w,\psi)}S_{(w,\varphi)} \in M^{(w,\varphi,\psi)}$ and (2) for all $\overrightarrow{w_n}, \overrightarrow{v_m} \in M^{(w,\varphi,\psi)}$, if $\overrightarrow{w_n} \in M^{(w,\varphi,\psi)}$ and either $\overrightarrow{w_n}R_{\Box}^{(w,\varphi)}\overrightarrow{v_m}$ or $\overrightarrow{w_n}R_{\mathsf{X}}^{(w,\varphi)}\overrightarrow{v_m}$, then $\overrightarrow{v_m} \in M^{(w,\varphi,\psi)}$.

The only relatively tricky part is 4.2 when $\blacksquare = \Box$. The proof is as follows: Let $\overrightarrow{w_n} = w_1 w_2 \ldots w_n \in W^{(w,\psi,\varphi)}$ and $\overrightarrow{u_k} = u_2 u_3 \ldots u_k \in W^{(w,\varphi)}$ be s.t. $f(\overrightarrow{w_n}) R_{\Box}^{(w,\varphi)} \overrightarrow{u_k}$. By the def. of $W^{(w,\psi,\varphi)}$, $w_1 R_{\mathsf{X}} w_2$. In addition, by the def. of $R_{\Box}^{(w,\varphi)}$, $w_2 R_{\Box} u_2$. Hence, by condition 3 from Def. 13, there is $v \in W$ s.t. $w_1 R_{\mathrm{Ag}} v$ and $v R_{\mathsf{X}} u_2$. It is not difficult to check that the sequence $v\overrightarrow{u_k}$ is s.t.: (1) $v\overrightarrow{u_k} \in M^{(w,\varphi,\psi)}$, (2) $\overrightarrow{w_n} R_{\Box}^{(w,\psi,\varphi)} v\overrightarrow{u_k}$, and (3) $f(v\overrightarrow{u_k}) = u_k$.

By standard results in modal logic [8, Prop. 2.14], fact 4 implies that

$$M^{(w,\psi,\varphi)}, \overrightarrow{w_n} \models \chi \text{ iff } M^{(w,\varphi)}, f(\overrightarrow{w_n}) \models \chi$$

for all $\overrightarrow{w_n} \in W^{(w,\psi,\varphi)}$ and $\chi \in \mathcal{L}'_{\mathsf{SLD}_n}$. Hence, $M^{(w,\psi,\varphi)}, \sigma_{(w,\psi)} \models \varphi$ iff $M^{(w,\varphi)}, \sigma_{(w,\varphi)} \models \varphi$ by fact 3, and so $M^{(w,\psi)}, \sigma_{(w,\psi)} \models \varphi$ iff $M^{(w,\varphi)}, \sigma_{(w,\varphi)} \models \varphi$ by fact 1. □

We are now ready to prove the central proposition of this part.

Proposition 6 *For any pointed pseudo-model M, w and $\varphi \in \mathcal{L}'_{\mathsf{SLD}_n}$, $M, w \models \varphi$ iff $M^{(w,\varphi)}, \sigma_{(w,\varphi)} \models \varphi$.*

Proof The proof is by $<_c^S$-induction on φ. The cases in which $\varphi := p$, $\varphi := do(a_i)$, and $\varphi := dev(a_i)$ follow immediately from Def. 21 and the fact that w is the last element of $\sigma_{w,p}$. For the inductive cases, we assume the following inductive hypothesis (IH): if M, v is a pointed pseudo-model and $\psi \in \mathcal{L}'_{\mathsf{SLD}_n}$ is s.t. $\psi <_c^S \varphi$, then $M, v \models \psi$ iff $M^{(v,\psi)}, \sigma_{(v,\psi)} \models \psi$. We omit the proof for the case in which $\varphi := \neg\psi$, which follows immediately from Remark 4. The other cases are as follows:

1. $\varphi := \psi \wedge \chi$.

   Suppose, without loss of generality, that $d(\psi) \leq d(\chi)$. Then, by Lem. 2, there is $\psi' \in \mathcal{L}'_{\mathsf{SLD}_n}$ s.t. (1) $\psi \leftrightarrow \psi'$ is valid in the class of pseudo-models, (2) $d(\psi') = d(\chi)$, and (3) $c(\psi') = c(\psi) < c(\psi) + c(\chi) + 1 = c(\psi \wedge \chi)$. It follows from (3) that (A) $\psi' <_c^S (\psi \wedge \chi)$ and $\chi <_c^S (\psi \wedge \chi)$ and from (2) that (B) $d(\psi \wedge \chi) = max\{d(\psi), d(\chi)\} = d(\chi) = d(\psi')$. Given these facts, we reason as follows:

$M, w \models \psi \wedge \chi$		
iff	$M, w \models \psi$ and $M, w \models \chi$	by def. of truth
iff	$M, w \models \psi'$ and $M, w \models \chi$	by (1)
iff	$M^{(w,\psi')}, \sigma_{(w,\psi')} \models \psi'$ and $M^{(w,\chi)}, \sigma_{(w,\chi)} \models \chi$	by IH, given (A)
iff	$M^{(w,\psi\wedge\chi)}, \sigma_{(w,\psi\wedge\chi)} \models \psi'$ and $M^{(w,\psi\wedge\chi)}, \sigma_{(w,\psi\wedge\chi)} \models \chi$	by Remark 4
iff	$M^{(w,\psi\wedge\chi)}, \sigma_{(w,\psi\wedge\chi)} \models \psi$ and $M^{(w,\psi\wedge\chi)}, \sigma_{(w,\psi\wedge\chi)} \models \chi$	by (1), given Prop. 5
iff	$M^{(w,\psi\wedge\chi)}, \sigma_{(w,\psi\wedge\chi)} \models \psi \wedge \chi$	by def. of truth

2. $\varphi := \Box\psi$.
Since $d(\psi) = d(\Box\psi)$, we can use Lemma 4:

$M, w \models \Box\psi$		
iff	for all v s.t. $wR_{\Box}v$, $M, v \models \psi$	by def. of truth
iff	for all v s.t. $wR_{\Box}v$, $M^{(v,\psi)}, \sigma_{(v,\psi)} \models \psi$	by IH
iff	for all v s.t. $\sigma_{(w,\Box\psi)}R_{\Box}^{(w,\Box\psi)}\sigma_{(v,\psi)}$, $M^{(w,\Box\psi)}, \sigma_{(v,\psi)} \models \psi$	by Lem. 4 and the def. of $R_{\Box}^{(w,\Box\psi)}$
iff	for all $\overrightarrow{v_n}$ s.t. $\sigma_{(w,\Box\psi)}R_{\Box}^{(w,\Box\psi)}\overrightarrow{v_n}$, $M^{(w,\Box\psi)}, \overrightarrow{v_n} \models \psi$	with $n = d(v,\psi)+1$
iff	$M^{(w,\Box\psi)}, \sigma_{(w,\Box\psi)} \models \Box\psi$	by def. of truth

3. $\varphi := \mathsf{X}\psi$.
For this case, we exploit the following facts: (A) $\psi \leftrightarrow \boxplus\psi$ is valid in the class of pseudo-models, (B) $\boxplus\psi <_c^S \mathsf{X}\psi$, since $c(\boxplus\psi) = c(\psi) = c(\mathsf{X}\psi)$ and $S(\boxplus\psi) = S(\psi)+1 < S(\psi)+2 = S(\mathsf{X}\psi)$.

$M, w \models \mathsf{X}\psi$		
iff	for all v s.t. $wR_{\mathsf{X}}v$, $M, v \models \psi$	by def. of truth
iff	for all v s.t. $wR_{\mathsf{X}}v$, $M, v \models \boxplus\psi$	by (A)
iff	for all v s.t. $wR_{\mathsf{X}}v$, $M^{(v,\boxplus\psi)}, \sigma_{(v,\boxplus\psi)} \models \boxplus\psi$	by IH, given (B)
iff	for all v s.t. $\sigma_{(w,\mathsf{X}\psi)}R_{\mathsf{X}}^{(w,\mathsf{X}\psi)}\sigma_{(v,\boxplus\psi)}$, $M^{(w,\mathsf{X}\psi)}, \sigma_{(v,\boxplus\psi)} \models \boxplus\psi$	by Lem. 5 and the def. of $R_{\mathsf{X}}^{(w,\mathsf{X}\psi)}$
iff	for all $\overrightarrow{v_n}$ s.t. $\sigma_{(w,\mathsf{X}\psi)}R_{\mathsf{X}}^{(w,\mathsf{X}\psi)}\overrightarrow{v_n}$, $M^{(w,\mathsf{X}\psi)}, \overrightarrow{v_n} \models \boxplus\psi$	with $n = d(v,\boxplus\psi)+1$
iff	for all $\overrightarrow{v_n}$ s.t. $\sigma_{(w,\mathsf{X}\psi)}R_{\mathsf{X}}^{(w,\mathsf{X}\psi)}\overrightarrow{v_n}$, $M^{(w,\mathsf{X}\psi)}, \overrightarrow{v_n} \models \psi$	by (A), given Prop. 5
iff	$M^{(w,\mathsf{X}\psi)}, \sigma_{(w,\mathsf{X}\psi)} \models \mathsf{X}\psi$	by def. of truth

The remaining cases are proved in a similar way. In particular, the case in which $\varphi := \mathsf{Y}\psi$ follows from Lemma 5 and the fact that R_{Y} is the converse R_{X}, while the case in which $\varphi := \boxplus\psi$ follows from Lemma 6 and the fact that $\boxplus\psi \leftrightarrow \psi$ is valid on any pseudo-model. □

Step 3: Dividing histories

We now want to show that a pseudo-model like $M^{(w,\varphi)}$ (where $\varphi \in \mathcal{L}_{\mathsf{SLD}_n}$) can be turned into an SLD_n model. The idea is simple: we take equivalence classes determined by $R_{\Box}^{(w,\varphi)}$ as moments and we show that $R_{\mathsf{X}}^{(w,\varphi)}$ induces a treelike ordering on moments. Before doing this, we take an extra step to ensure that the states in $W^{(w,\varphi)}$ and the moment-history pairs in the resulting SLD_n model will be in a one-to-one correspondence.

Definition 22 (X-*depth* of $\varphi \in \mathcal{L}_{\mathsf{SLD}_n}$) The X-depth $x(\varphi)$ of $\varphi \in \mathcal{L}_{\mathsf{SLD}_n}$ is defined as:

$$\begin{aligned} x(p) = x(do(a_i)) = x(dev(a_i)) &= 0 \\ x(\neg\varphi) = x(\Box\varphi) = x(\mathsf{Y}\varphi) &= x(\varphi) \\ x(\mathsf{X}\varphi) &= x(\varphi)+1 \end{aligned}$$

Definition 23 Let M, w be a pointed pseudo-model and $\varphi \in \mathcal{L}_{\mathsf{SLD}_n}$. Then, $M^{(w,\varphi,x)}$ is the tuple obtained from $M^{(w,\varphi)}$ by replacing $R_{\Box}^{(w,\varphi)}$ with the relation $R_{\Box}^{(w,\varphi,x)}$ defined by setting, for all $\overrightarrow{w_n}, \overrightarrow{v_m} \in W^{(w,\varphi)}$,

$$\overrightarrow{w_n} R_{\Box}^{(w,\varphi,x)} \overrightarrow{v_m} \text{ iff } \begin{cases} \overrightarrow{w_n} R_{\Box}^{(w,\varphi)} \overrightarrow{v_m} & \text{if } n \leq d(w,\varphi) + x(\varphi) + 1 \\ \overrightarrow{w_n} = \overrightarrow{v_m} & \text{otherwise} \end{cases}$$

So, in $M^{(w,\varphi,x)}$, all sequences of length $n > d(w,\varphi) + x(\varphi) + 1$ belong to a singleton equivalence class of $R_{\Box}^{(w,\varphi,x)}$. It is immediate to check that $M^{(w,\varphi,x)}$ is still a pseudo-model. In addition, the next proposition follows straightforwardly from Proposition 5 and from the fact that $R_{\Box}^{(w,\varphi)}$-equivalent states are separated in $M^{(w,\varphi,x)}$ by taking into account the modal X-depth of φ.

Proposition 7 *For any pointed pseudo-model M, w and $\varphi \in \mathcal{L}_{SLD_n}$, $M, w \models \varphi$ iff $M^{(w,\varphi,x)}, \sigma_{(w,\varphi)} \models \varphi$.*

Step 4: From pseudo-models to SLD_n models

Let $\varphi \in \mathcal{L}_{\mathsf{SLD}_n}$ be an SLD_n-consistent formula and M, w a pointed pseudo-model s.t. $M, w \models \varphi$. Then, $M^{(w,\varphi,x)}, \sigma_{(w,\varphi)} \models \varphi$ by Prop. 7. Define $\mathcal{T} = \langle Mom, m_0, < \rangle$ so that:

- Mom is the quotient set of $W^{(w,\varphi)}$ by $R_{\Box}^{(w,\varphi,x)}$;
- $m_0 = [s_{(w,\varphi)}]$ is the equivalence class in Mom of the one-element sequence $s_{(w,\varphi)}$.
- $< \subseteq Mom \times Mom$ is s.t., for all $[\overrightarrow{w_n}], [\overrightarrow{v_m}] \in Mom$, $[\overrightarrow{w_n}] < [\overrightarrow{v_m}]$ iff all prefixes of length n of sequences in $[\overrightarrow{v_m}]$ are in $[\overrightarrow{w_n}]$ (*i.e.*, iff $n < m$ and, for all $\overrightarrow{u_m} \in [\overrightarrow{v_m}]$, $\overrightarrow{u_n} \in [\overrightarrow{w_n}]$).

It is not difficult to check that $\mathcal{T}$ is a DBT structure. In addition, there is a one-to-one correspondence between possible states in the pseudo-model $M^{(w,\varphi,x)}$ and indices in $\mathcal{T}$. To see this, where $n \in \mathbb{N}$ and $h \in Hist^{\mathcal{T}}$, let $h(n)$ be the n-th moment on h.[34] Any index $[\overrightarrow{w_n}]/h$ can then be re-written as $h(n)/h$. Finally, let $\mathsf{z} = d(w,\varphi) + x(\varphi) + 1$. Observe that Definition 23 and the functionality of $R_{\mathsf{X}}^{(w,\varphi)}$ ensure that, for all $n > \mathsf{z}$, $h(n)$ is a singleton (we write $\overrightarrow{w_{h(n)}}$ for its only element). Define a mapping $\omega : Ind^{\mathcal{T}} \to W^{(w,\varphi)}$ by setting: for all $h(n)/h \in Ind^{\mathcal{T}}$,

$$\omega(h(n)/h) \text{ is the prefix of length n of } \overrightarrow{w_{h(n+\mathsf{z})}}.$$

Intuitively, the function ω finds, for every index $h(n)/h$, the "witness" of h in $h(n)$. It does so by picking a singleton moment on h that occurs later than $h(n)$ and by

[34] Formally, $h(n)$ is defined inductively as follows: $h(0) = m_0$; $h(n+1) = succ_h(h(n))$.

selecting the prefix of length n of its unique element.[35] It is an easy exercise to prove that ω is a bijection. We write ω^{-1} for its inverse.

Now, where $\mathcal{T}$ is defined as dabove, define $\mathcal{M} = \langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \pi \rangle$ so that:

- $\mathbf{act} : Ind^{\mathcal{T}} \to Ag\text{-}Acts$ is s.t. $\mathbf{act}(h(n)/h) = f_{do}^{(w,\varphi)}(\omega(h(n)/h))$;
- $\mathbf{dev} : Mom \to 2^{Acts}$ is s.t. $\mathbf{dev}([\overrightarrow{w_n}]) = f_{dev}^{(w,\varphi)}(\overrightarrow{w_n})$;
- $\pi : Prop \to 2^{Ind^{\mathcal{T}}}$ is s.t. $h(n)/h \in \pi(p)$ iff $\omega(h(n)/h) \in v^{(w,\varphi)}(p)$.

Proposition 8 *$\mathcal{M}$ is a SLD_n model.*

Proof We noticed above that $\mathcal{T}$ is a DBT structure. It is immediate to see that the function **dev** is well defined because $f_{dev}^{(w,\varphi)}$ satisfies condition 4.1 from Def. 13. In addition, **dev** also satisfies conditions 3, 4, and 5 from Def. 4 because $f_{dev}^{(w,\varphi)}$ satisfies the corresponding conditions 4.2, 4.3, and 4.4 from Def. 13. For the remaining conditions the proof can be easily adapted from the proof of Proposition A.1.23 in [18, p. 211]. □

Proposition 9 *For all formulas $\psi \in \mathcal{L}_{\mathsf{SLD}_n}$ and indices $h(n)/h$ in $\mathcal{M}$,*

$$\mathcal{M}, h(n)/h \models \psi \text{ iff } M^{(w,\varphi,x)}, \omega(h(n)/h) \models \psi$$

Proof The proof is by induction on the complexity of ψ. The cases for propositional variables, Boolean connectives, and formulas like $do(a_i)$ and $dev(a_i)$ follow straightforwardly from the def. of $\mathcal{M}$. The proof for the cases in which $\psi := \Box\chi$ and $\psi := \mathsf{X}\chi$ can be easily adapted from the proof of Proposition A.1.24 in [18, pp. 211-212]. Finally, the case in which $\psi := \mathsf{Y}\chi$ is analogous to the case in which $\psi := \mathsf{X}\chi$. □

Proposition 9 and the fact that $M^{(w,\varphi,x)}, \sigma_{(w,\varphi)} \models \varphi$ entail that $\mathcal{M}, \omega^{-1}(\sigma_{(w,\varphi)}) \models \varphi$. Since φ is an arbitrary SLD_n-consistent formula and $\mathcal{M}$ is an SLD_n model, we can then conclude that SLD_n is complete w.r.t. the class of all SLD_n models.

B Proofs of Propositions 3 and 4

Proof of Proposition 3 We only present the proof for the left-to-right direction of $\mathsf{Dis}_{\mathsf{X}}$ and for $\mathsf{Cen1}$, as the remaining cases are similar. Let $\mathcal{M} = \langle \langle Mom, m_0, < \rangle, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$ be either a rewind model or an independence model and m/h an index in $\mathcal{M}$. Recall that, for any $h \in H_m$, $succ_h(m)$ is the successor of m on history h and that t_m is the instant to which m belongs. Definition 2 ensures that, for any $h \in H_m$ and $h' \in H_{m'}$, $\mathsf{t}_m = \mathsf{t}_{m'}$ iff $\mathsf{t}_{succ_h(m)} = \mathsf{t}_{succ_{h'}(m')}$. Below, we will repeatedly use this fact without explicit mention.

[35] The prefix of length n of $\overrightarrow{w_{h(n+z)}}$ is an element of $h(n)$ by the definition of $<$.

(Dis$_{\mathsf{X}}$, L-R) If $\mathcal{M}, m/h \models \mathsf{X}(\varphi \,\Box\!\!\rightarrow \psi)$, then $\mathcal{M}, succ_h(m)/h \models \varphi \,\Box\!\!\rightarrow \psi$. There are two cases. **Case 1:** There is no $h' \in Hist$ s.t. $\mathcal{M}, \mathrm{t}_{succ_h(m)}/h' \models \varphi$. Then, there is no h' s.t. $\mathcal{M}, \mathrm{t}_m/h' \models \mathsf{X}\varphi$, otherwise $\mathcal{M}, \mathrm{t}_{succ_h(m)}/h' \models \varphi$, against the hypothesis. Hence, $\mathcal{M}, m/h \models \mathsf{X}\varphi \,\Box\!\!\rightarrow \psi \mathsf{X}\psi$ by Def. 8 (i). **Case 2:** There is $h' \in Hist$ s.t. $\mathcal{M}, \mathrm{t}_{succ_h(m)}/h' \models \varphi \wedge \psi$ and, for all $h'' \in Hist$, if $\mathcal{M}, \mathrm{t}_{succ_h(m)}/h'' \models \varphi \wedge \neg\psi$, then $h'' \not\simeq_h h'$. If $\mathcal{M}, \mathrm{t}_{succ_h(m)}/h' \models \varphi \wedge \psi$, then $\mathcal{M}, \mathrm{t}_m/h' \models \mathsf{X}\varphi \wedge \mathsf{X}\psi$. Take any $h^* \in Hist$ s.t. $(*)$ $\mathcal{M}, \mathrm{t}_m/h^* \models \mathsf{X}\varphi \wedge \neg\mathsf{X}\psi$. We want to show that $h^* \not\simeq_h h'$. By the def. of truth, $(*)$ implies that $\mathcal{M}, \mathrm{t}_{succ_h(m)}/h^* \models \varphi \wedge \neg\psi$, and so $h^* \not\simeq_h h'$ by our hypothesis. Hence, $\mathcal{M}, m/h \models \mathsf{X}\varphi \,\Box\!\!\rightarrow \psi \mathsf{X}\psi$ by Definition 8 (ii).

(Cen1) Assume that $\mathcal{M}, m/h \models \Diamond\varphi \wedge \Diamond\psi$. Then, $(*)$ there is $h' \in H_m$ s.t. $\mathcal{M}, m/h' \models \varphi \wedge \Diamond\psi$. Take any $h'' \in H_m$. We want to show that $\mathcal{M}, m/h'' \models \varphi \,\Box\!\!\rightarrow \psi \Diamond\psi$. Given $(*)$, the vacuous case is excluded. So, consider any history h^* s.t. (2) $\mathcal{M}, \mathrm{t}_m/h^* \models \varphi \wedge \neg\Diamond\psi$. We want to show that $h^* \not\simeq_{h''} h'$. Since $\mathcal{M}, m/h'' \models \Diamond\psi$ and $\mathcal{M}, \mathrm{t}_m/h^* \not\models \Diamond\psi$, $h^* \notin H_m$, *i.e.*, h^* branches off from h'' earlier than m. Since $h' \in H_m$, this means that $past\_ov(h'', h') \supset past\_ov(h'', h^*)$, and so (3) $h^* \not\simeq_{h''} h'$ by Def. 9. Since h^* is an arbitrary history satisfying (2), (1) and (3) suffice to conclude that $\mathcal{M}, m/h'' \models \varphi \,\Box\!\!\rightarrow \psi$. Hence, $\mathcal{M}, m/h \models \Box(\varphi \,\Box\!\!\rightarrow \psi)$, as h'' is an arbitrary history in H_m.

Proof of Proposition 4: Part 1 To see that Exp$_{\Box}$ is valid on any rewind model, let $\mathcal{M} = \left\langle \langle Mom, m_0, < \rangle, \mathbf{act}, \mathbf{dev}, \preceq^R \right\rangle$ be a rewind model and m/h any index in $\mathcal{M}$. Assume that $\mathcal{M}, m/h \models \Box\neg\varphi \wedge (\varphi \,\Box\!\!\rightarrow \psi \Box\psi)$. We have to show that $\mathcal{M}, m/h \models \Box(\varphi \,\Box\!\!\rightarrow \psi\psi)$. Since $\mathcal{M}, m/h \models \varphi \,\Box\!\!\rightarrow \psi\Box\psi$ by hypothesis, there are two cases. **Case 1:** There is no $h' \in Hist$ s.t. $\mathcal{M}, \mathrm{t}_m/h' \models \varphi$. Then, for any $h'' \in H_m$, $\mathcal{M}, m/h'' \models \varphi \,\Box\!\!\rightarrow \psi$ by Def. 8 (i). Hence, $\mathcal{M}, m/h \models \Box(\varphi \,\Box\!\!\rightarrow \psi\psi)$ by Def. 7. **Case 2:** There is $h' \in Hist$ s.t. (1) $\mathcal{M}, \mathrm{t}_m/h' \models \varphi \wedge \Box\psi$ and (2) for all $h'' \in Hist$ s.t. $\mathcal{M}, \mathrm{t}_m/h'' \models \varphi \wedge \neg\Box\psi$, $h'' \not\preceq^R_h h'$. Take any $h^* \in H_m$. We want to show that $\mathcal{M}, m/h^* \models \varphi \,\Box\!\!\rightarrow \psi\psi$. By (1), we know that (3) $\mathcal{M}, \mathrm{t}_m/h' \models \varphi \wedge \psi$. So, consider any $h''' \in Hist$ s.t. (4) $\mathcal{M}, \mathrm{t}_m/h''' \models \varphi \wedge \neg\psi$. We want to prove that $h''' \not\preceq^R_{h^*} h'$, i.e., that $h' \prec^R_{h^*} h'''$ and $h''' \not\prec^R_{h^*} h'$. In order to prove that $h' \prec^R_{h^*} h'''$ we need to prove the following:

$$
\begin{array}{l}
past\_ov(h^*, h') \supset past\_ov(h^*, h'''), \ or \\
past\_ov(h^*, h') = past\_ov(h^*, h''') \ \text{ and } n\_sep(h^*, h') < n\_sep(h^*, h'''), \ or \\
past\_ov(h^*, h') = past\_ov(h^*, h''') \ \text{ and } n\_sep(h^*, h') = n\_sep(h^*, h''') \ \text{ and} \\
n\_dev(h''') < n\_dev(h').
\end{array}
$$

Observe that:

(a) $\mathcal{M}, \mathrm{t}_m/h''' \models \varphi \wedge \neg\Box\psi$ by (4), and so $h''' \not\preceq^R_h h'$ by (2). By the definition of $\preceq^R_h$, it follows that $h' \prec^R_h h'''$, i.e., that

$$
\begin{array}{l}
past\_ov(h, h') \supset past\_ov(h, h'''), \ or \\
past\_ov(h, h') = past\_ov(h, h''') \ \text{ and } n\_sep(h, h') < n\_sep(h, h'''), \ or \\
past\_ov(h, h') = past\_ov(h, h''') \ \text{ and } n\_sep(h, h') = n\_sep(h, h''') \ \text{ and} \\
n\_dev(h''') < n\_dev(h').
\end{array}
$$

(b) Since $h, h^* \in H_m$, for all $m' \leq m$, the initial segment of h up to m' and the initial segment of h^* up to m' are equal.

(c) $past\_ov(h, h') = past\_ov(h^*, h')$. In fact, $\mathcal{M}, m/h \models \Box\neg\varphi$ by hypothesis, while $\mathcal{M}, \mathsf{t}_m/h' \not\models \Box\neg\varphi$ by (1). Hence, h' must branch off from h at some moment $m' < m$. Since, by (b), the initial segment of h up to m' is the same as the initial segment of h^* up to m', h' must also branch off from h^* at m'. Thus, $h \cap h' = h^* \cap h'$.

(d) $past\_ov(h, h''') = past\_ov(h^*, h''')$: analogous to (c).

(e) $num\_sep(h, h') = num\_sep(h^*, h')$. In fact, by (b), h and h^* are undivided at all moments $m'' < m$. By the condition of no choice between undivided histories, this means that, for all such m'', $\mathbf{act}(m''/h) = \mathbf{act}(m''/h^*)$. Now, as we have seen in item (c), there is a moment $m' < m$ s.t. h' branches off from both h and h^* at m'. Since $\mathbf{act}(m'/h) = \mathbf{act}(m'/h^*)$, it follows that, for all $i \in Ag$, $\mathbf{act}(m'/h)(i) \neq \mathbf{act}(m'/h')(i)$ iff $\mathbf{act}(m'/h^*)(i) \neq \mathbf{act}(m'/h')(i)$. Hence, $|\{i \in Ag \,|\, \mathbf{act}(m'/h)(i) \neq \mathbf{act}(m'/h')(i)\}| = |\{i \in Ag \,|\, \mathbf{act}(m'/h^*)(i) \neq \mathbf{act}(m'/h')(i)\}|$.

(f) $num\_sep(h, h''') = num\_sep(h^*, h''')$: analogous to (e).

From items (c) to (f) it follows that $h' \prec^R_h h'''$ iff $h' \prec^R_{h^*} h'''$; from this and item (a) it follows that $h' \prec^R_{h^*} h'''$. The proof that $h''' \not\prec^R_{h^*} h'$ proceeds in a similar way.

Proof of Proposition 4: Part 2 To see that $\mathsf{Exp}_\Box$ is invalid in some independence model, consider Fig. 7. Assume that: (1) for any agents $i \in Ag \setminus \{1\}$ and moment m', $Acts_i^{m'} = \{vc_i\}$, (2) for any moment m' not depicted in the figure $Acts_1^{m'} = \{vc_1\}$, and (3) for any moment m', $\mathbf{dev}(m') = \varnothing$. It is not difficult to check that the defined structure is an SLD_n frame. As shown in the figure, let p be true at t_2/h_5, t_2/h_6, t_2/h_{11}, t_2/h_{12} and q be true at t_2/h_5, t_2/h_6, t_2/h_7, and t_2/h_8. Then, $\mathsf{t}_2/h_1 \models \Box\neg p$ and $\mathsf{t}_2/h_1 \models p \,\Box\!\!\rightarrow \Box q$. In fact, (1) the most similar history to h_1 where p is true at time t_2 is h_5, as all unconstrained agents (*i.e.*, all agents) do the same types of action on h_1 and h_5 at all times, and (2) $\Box q$ is true at t_2/h_5. On the other hand, $\mathsf{t}_2/h_1 \not\models \Box(p \,\Box\!\!\rightarrow q)$. Consider, in fact, history h_3: The most similar history to h_3 where p is true at time t_2 is h_{11}, as all unconstrained agents do the same types of

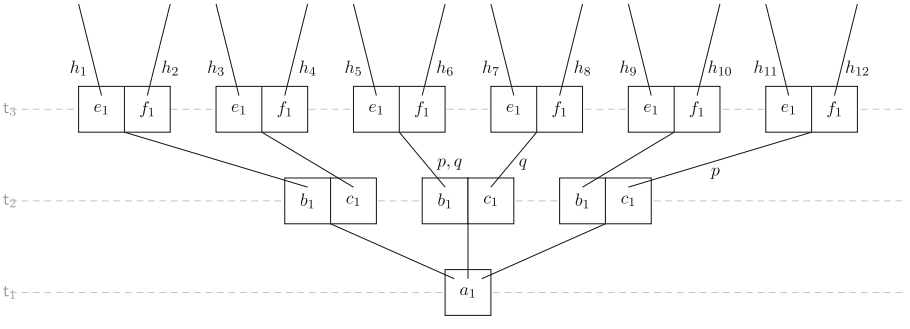

Fig. 7 An independence model not satisfying $\mathsf{Exp}_\Box$

action on h_3 and h_{11} at all times. Since q is false at t_2/h_{11}, $t_2/h_3 \not\models p \Box\!\!\rightarrow q$. Therefore, $t_2/h_1 \not\models \Box(p \Box\!\!\rightarrow q)$.

Remark 5 The model depicted in Fig. 7 satisfies the conditions of uniformity of menus and of identity of overlapping menus from Section 3.2. Hence, $\mathsf{Exp}_{\Box}$ remains invalid in the class of independence models satisfying these conditions.

Acknowledgements We would like to thank the participants of the audience of the following seminars and workshop at which this paper was presented: Logic Seminar (University of Maryland), LIRa Seminar (University of Amsterdam, 2020), Tsinghua Online Logic Seminar (Tsinghua University, 2020), LACN-Workshop (University of Amsterdam, 2020), Ghent-Brussels Seminar (University of Ghent and University of Brussels, 2021). We would also like to thank Paolo Santorio and two anonymous referees for their valuable comments.

References

1. Aumann, R. (1995). Backward induction and common knowledge of rationality. *Games Econ. Behav.*, *8*(1), 6–19.
2. Baltag, A., Canavotto, I., & Smets, S. (2021). Causal Agency and Responsibility: A Refinement of STIT Logic. In A. Giordani, & J. Malinowski (Eds.) *Logic in High Definition, Trends in Logical Semantics, volume 56 of Trends in Logic* (pp. 149–176). Berlin: Springer.
3. Battigalli, P. (1997). On rationalizability in extensive games. *J. Econ. Theory*, *74*, 40–61.
4. Battigalli, P., & Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *J. Econ. Theory*, *106*(2), 356–391.
5. Belnap, N., Perloff, M., & Xu, M. (2001). Facing the future: Agents and choices in our indeterministic world. Oxford University Press, Oxford.
6. Bennett, J. (2003). A Philosophical Guide to Conditionals. Clarendon Press, Oxford.
7. Bicchieri, C. (1988). Strategic behavior and counterfactuals. *Synthese*, *76*, 135–169.
8. Blackburn, P., de Rijke, M., & Yde, V. (2001). Modal Logic. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge.
9. Bonanno, G. (2014). A doxastic behavioral characterization of generalized backward induction. *Games Econ. Behav.*, *88*, 221–241.
10. Bonanno, G. (2015). Counterfactuals and the Prisoner's Dilemma. In The Prisoner's Dilemma, pp. 133–155. Cambridge University Press.
11. Broersen, J. M. (2011a). Deontic Epistemic stit Logic Distinguishing Modes of Mens Rea. *J. Appl. Log.*, *9*(2), 137–152.
12. Broersen, J. M. (2011b). Making a Start with the stit Logic Analysis of Intentional Action. *J. Philos. Log.*, *40*(4), 499–530.
13. Broersen, J. M. (2013). Probabilistic stit Logic and its Decomposition. *Int. J. Approx. Reason.*, *54*, 467–477.
14. Broersen, J. M. (2014). On the Reconciliation of Logics of Agency and Logics of Event Types. In R. Trypuz (Ed.) *Krister Segerberg on Logic of Actions, volume 1 of Outstanding Contributions to Logic* (pp. 41–59). Netherlands: Springer.

15. Broersen, J. M., & Herzig, A. (2015). Using STIT Theory to Talk About Strategies. In J. Benthem, S. Ghosh, & S. Verbrugge (Eds.) *Models of Strategic Reasoning. Logics, Games, and Communities* (pp. 137–173). Berlin: Springer.
16. Broersen, J. M., Herzig, A., & Troquard, N. (2006). From coalition logic to STIT. *Electron. Notes Theor. Comput. Sci.*, *157*(4), 23–35.
17. Broersen, J. M., & Ramírez Abarca, A. I. (2018). Knowledge and Subjective Oughts in STIT Logic. In J. M. Broersen, C. Condoravdi, S. Nair, & G. Pigozzi (Eds.) *Deontic Logic and Normative Systems, 14th International Conference (DEON 2018)* (pp. 51–69). Milton Keynes: College Publications.
18. Canavotto, I. (2020). Where Resposibility Takes You. Logics of Agency, Counterfactuals and Norms. PhD thesis, Institute for logic, Language and Computation. University of Amsterdam.
19. Ciuni, R., & Horty, J. F. (2014). Stit Logics, Games, Knowledge, and Freedom. In A. Baltag, S. Smets, & J. van Benthem (Eds.) *on Logic and Information Dynamics, volume 5 of Outstanding Contributions to Logic* (pp. 631–656). Cham: Springer.
20. Ciuni, R., & Mastop, R. (2009). Attributing Distributed Responsibility in Stit Logic. In X. He, J. F. Horty, & E. Pacuit (Eds.) *Logic, Rationality, and Interaction* (pp. 66–75). Berlin: Springer.
21. Dietrich, F., & List, C. (2016). Reason-based Choice and Context-dependence: An explanatory framework. *Econ. Philos.*, *2*(32), 175–229.
22. Harel, D., Kozen, D., & Jerzy, T. (2000). Dynamic Logic. The MIT Press, Cambridge.
23. Herzig, A., & Troquard, N. (2006). Knowing how to play: Uniform choices in logics of agency. In: Proceedings of the 5th International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS-06), pp. 209–216. The Association for Computing Machinery Press, New York.
24. Herzig, A., & Lorini, E. (2010). A dynamic logic of agency I: STIT, capabilities and Powers. *J. Log. Lang. Inf.*, *19*(1), 89–121.
25. Horty, J. F. (2001). Agency and Deontic Logic. Oxford University Press, Oxford.
26. Horty, J. F. (2012). Reasons as Defaults. Oxford University Press, Oxford.
27. Horty, J. F., & Pacuit, E. (2017). Action types in stit semantics. *Rev. Symbol. Log.*, *10*(4), 617–637.
28. Kalai, G., Rubinstein, A., & Spiegler, R. (2002). Rationalizing choice functions by mutltiple rationales. *Econometrica*, *70*(6), 2481–2488.
29. Kooi, B., & Tamminga, A. (2008). Moral conflicts between groups of agents. *J. Philos. Log.*, *37*(1), 1–21.
30. Lewis, D. (1973). Counterfactuals. Harvard University Press, Cambridge.
31. Lewis, D. (1979). Counterfactual dependence and time's arrow. *Nous*, *13*(4), 455–476.
32. Lorini, E., & Longin, D. (2014). Eunate mayor. A logical analysis of responsibility attribution: Emotions, individuals and collectives. *J. Log. Comput.*, *24*(6), 1313–1339.
33. Lorini, E., & Sartor, G. (2016). A STIT logic for reasoning about social influence. *Studia Log.*, *104*(4), 773–812.
34. Müller, T. (2005). On the Formal Structure of Continuous Action. In R. Schmidt, I. Pratt-Hartmann, M. Reynolds, & H. Wansing (Eds.) *Advances in Modal Logic*, (Vol. 5 pp. 191–209). London: King's College Publications.
35. Pauly, M. (2002). A modal logic for coalitional power in games. *J. Log. Comput.*, *12*(1), 149–166.
36. Pearl, J. (2000). Causality. Models, Reasoning, and Inference. Cambridge University Press, Cambridge.
37. Perea, A. (2014). Belief in the opponents' future rationality. *Games Econ. Behav.*, *83*, 231–254.
38. Placek, T., & Müller, T. (2007). Counterfactuals and historical possibility. *Synthese*, *154*(2), 173–197.
39. Selten, R., & Leopold, U. (1982). Subjunctive Conditionals in Decision and Game Theory. In Philosophy of Economics, pp. 191–200. Springer.
40. Sen, A. (1997). Maximization and the act of choice. *Econometrica*, *65*(4), 745–779.
41. Shin, H. S. (1992). Counterfactuals and a Theory of Equilibrium in Games. In C. Bicchieri, & M. L. D. Chiara (Eds.) *Knowledge, Belief, and Strategic Interaction* (pp. 397–413). Cambridge: Cambridge University Press.
42. Shoham, Y. (1989). Time for Action: On the Relation between Time, Knowledge and Action. In: Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI'89), vol. 2, pp. 954–959. Morgan Kaufmann Publishers, San Francisco.
43. Skyrms, B. (1998). Bayesian Subjunctive Conditionals for Games and Decisions. In Game Theory, Experience, Rationality, pp. 161–172.
44. Slote, M. A. (1978). Time in counterfactuals. *Philos. Rev.*, *87*(1), 3–27.

45. Stalnaker, R. (1996). Knowledge: Belief and counterfactual reasoning in games. *Econ. Philos.*, *12*, 133–163.
46. Stalnaker, R. C. (1968). A Theory of Conditionals. In R. Nicholas (Ed.) *Studies in Logical Theory* (pp. 98–112). Oxford: Basil Blackwell.
47. Stalnaker, R. C. (1998). Belief revision in games: Forward and backward induction. *Math. Soc. Sci.*, *36*(1), 31–56.
48. Tamminga, S. (2013). Deontic logic for strategic games. *Erkenntnis*, *78*(1), 183–200.
49. Thomason, R. H., & Gupta, A. (1981). A Theory of Conditionals in the Context of Branching Time. In W. L. Harper, R. C. Stalnaker, & G. Pearce (Eds.) *IFS: Conditionals, Belief, Decision, Chance and Time* (pp. 299–322). Netherlands: Springer.
50. Troquard, N., & Vieu, L. (2006). Towards a Logic of Agency and Actions with Duration. In: European Conference on Artificial Intelligence 2006 (ECAI'06), volume 141 of Frontiers in Artificial Intelligence and Applications, pp. 775–776. IOS Press, Amsterdam.
51. Turrini, P. (2012). Agreements as Norms. In T. Ågotnes, J. Broersen, & D. Elgesem (Eds.) *Deontic Logic in Computer Science, 11th International Conference (DEON 2011)* (pp. 31–45). Berlin: Springer.
52. Xu, M. (1997). Causation in branching time (I): Transitions, events and causes. *Synthese*, *112*(2), 137–192.
53. Xu, M. (2010). Combinations of Stit and Actions. *J. Log. Lang. Inf.*, *19*(4), 485–503.
54. Zambrano, E. (ed.) (2004). Counterfactual Reasoning and Common Knowledge of Rationality in Normal Form Games. *Top. Theor. Econ.* **4**(8).