# Conditionals in Game Theory

Ilaria Canavotto, University of Maryland
Eric Pacuit, University of Maryland

Lecture 5

ESSLLI 2022

# What we have done so far

1. Introduction to game theory
2. Conditionals in normal form games:
   - 2.1 Counterfactual rationality and independence
   - 2.2 Conditional choice rules and communication
   - 2.3 Translucency
3. Conditionals in extensive form games
   - 3.1 Backward induction
   - 3.2 Forward induction

# Plan for today

4. More on conditionals in sequential choice problems

   4.1 Planning conditionals

S.M. Hutteger and G.J. Rothfus. *Bradley conditionals and dynamic choice*. Synthese 199: 6585-6599, 2021.

R. Bradley. *Decision theory with a human face*. Cambridge University Press, ch8, 2017.

   4.1 Choice driven counterfactuals in branching time

IC & Eric Pacuit. *Choice driven counterfactuals*. JPL, 51, pp. 297–345, 2022.

# Conditionals in planning

S.M. Hutteger and G.J. Rothfus. *Bradley conditionals and dynamic choice*. Synthese 199: 6585-6599, 2021.

*Most decision problems discussed in philosophy have a static flavor: an agent makes a one-time choice from among a set of acts. Many decision situations involve a temporal component, however. Choices are made sequentially, perhaps mixed with receiving partial information about the state of the world. How should an agent's actions be modeled to fit the sequential environment? (p. 6586)*

# Central notion: plan

- A plan is a course of action that extends over time
- A plan involves conditionals: it tells us what to do **if** an event happens for a range of events that we can anticipate *ex ante*

  *Example*: I plan to fly to Dulles International Airport *and then* take the metro to DC **if** my flight is on time *and* take an Uber **if** my flight is delayed.
- A plan involves stability over time:

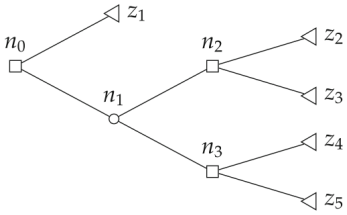
  *[I]f our initial endorsement of plan A over plan B is to be genuine, it presumably factors in all the contingencies we are aware of; thus, our endorsement "plan A is better than plan B" should not change at our whim. [p.6586]*

  - Counterpart in rational choice theory: dynamic consistency

# Main problem and main result

- According to rational choice theory, agents should be desirability maximizers: they should always select options with the highest desirability value.
- So, when options are plans, agents should select, at any moment of time, the plan with the highest desirability value.
- Key question: does selecting the plan with the highest desirability value preserve dynamic consistency?
- Hutteger's and Rothfus's answer: yes, if plans are understood as involving indicative conditionals

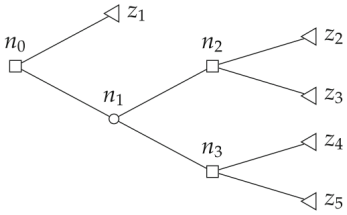
# Decision trees and plans




- □ is a choice node
- ○ is a nature node
- ◁ is a terminal node
- *move*$(n)$ says "the chooser moves to $n$"
- *stay*$(z)$ says "the agent stays at $z$"

A plan at a node of a tree specifies a unique move for every node *that the agent could reach*, given execution of earlier portions of the plan

# Decision trees and plans

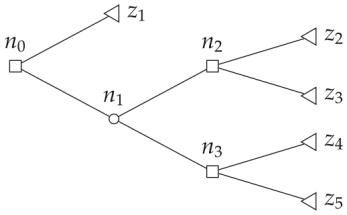



- □ is a choice node
- ○ is a nature node
- ◁ is a terminal node
- *move*$(n)$ says "the chooser moves to $n$"
- *stay*$(z)$ says "the agent stays at $z$"

A plan at a node of a tree specifies a unique move for every node *that the agent could reach*, given execution of earlier portions of the plan

Plans at $z_i$:
*stay*$(z_i)$

# Decision trees and plans




□ is a choice node

○ is a nature node

◁ is a terminal node

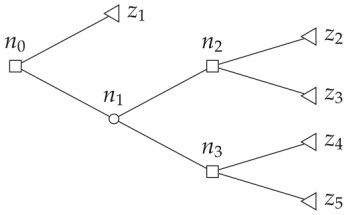
$move(n)$ says "the chooser moves to $n$"

$stay(z)$ says "the agent stays at $z$"

A plan at a node of a tree specifies a unique move for every node *that the agent could reach*, given execution of earlier portions of the plan

Plans at $n_2$:
$move(z_2) \wedge stay(z_2)$
$move(z_3) \wedge stay(z_3)$

# Decision trees and plans

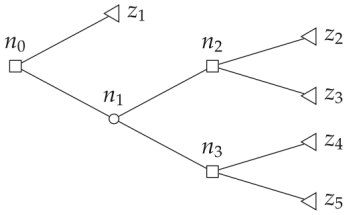



- $\square$ is a choice node
- $\bigcirc$ is a nature node
- $\triangleleft$ is a terminal node
- $move(n)$ says "the chooser moves to $n$"
- $stay(z)$ says "the agent stays at $z$"

A plan at a node of a tree specifies a unique move for every node *that the agent could reach*, given execution of earlier portions of the plan

Plans at $n_3$:
$move(z_4) \wedge stay(z_4)$
$move(z_5) \wedge stay(z_5)$

# Decision trees and plans



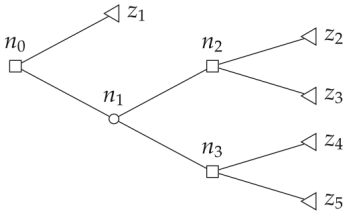

- $\square$ is a choice node
- $\bigcirc$ is a nature node
- $\triangleleft$ is a terminal node
- $move(n)$ says "the chooser moves to $n$"
- $stay(z)$ says "the agent stays at $z$"

A plan at a node of a tree specifies a unique move for every node *that the agent could reach*, given execution of earlier portions of the plan

Plans at $n_1$:

$(move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4)))$
$(move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_5) \wedge stay(z_5)))$
$(move(n_2) \rightarrow (move(z_3) \wedge stay(z_3))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4)))$
$(move(n_2) \rightarrow (move(z_3) \wedge stay(z_3))) \wedge (move(n_3) \rightarrow (move(z_5) \wedge stay(z_5)))$

# Decision trees and plans




- $\square$ is a choice node
- $\bigcirc$ is a nature node
- $\triangleleft$ is a terminal node
- $move(n)$ says "the chooser moves to $n$"
- $stay(z)$ says "the agent stays at $z$"

A plan at a node of a tree specifies a unique move for every node *that the agent could reach*, given execution of earlier portions of the plan

Plans at $n_0$:

$move(z_1) \wedge stay(z_1)$

$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4))))$

$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_5) \wedge stay(z_5))))$

$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_3) \wedge stay(z_3))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4))))$

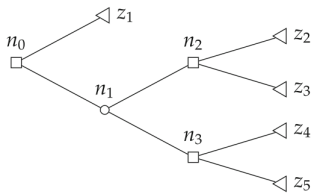
$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_3) \wedge stay(z_3))) \wedge (move(n_3) \rightarrow (move(z_5) \wedge stay(z_5))))$

# Continuation of a plan

If $\pi$ is a plan that makes arrival at $n$ possible, $\pi(n)$ **is the continuation of** $\pi$ **at** $n$

# Continuation of a plan

If $\pi$ is a plan that makes arrival at $n$ possible, $\pi(n)$ **is the continuation of** $\pi$ **at** $n$



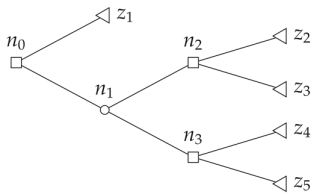

Let $\pi$ be the following plan available at $n_0$:

$$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4))))$$

Then:

# Continuation of a plan

If $\pi$ is a plan that makes arrival at $n$ possible, $\pi(n)$ **is the continuation of** $\pi$ **at** $n$




Let $\pi$ be the following plan available at $n_0$:

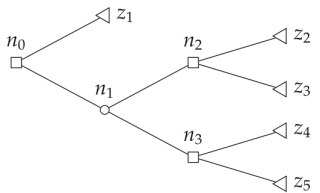
$$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4))))$$

Then:

- $\pi(n_1)$

# Continuation of a plan

If $\pi$ is a plan that makes arrival at $n$ possible, $\pi(n)$ **is the continuation of** $\pi$ **at** $n$




Let $\pi$ be the following plan available at $n_0$:

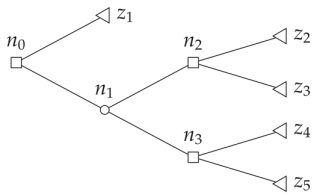
$$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4))))$$

Then:

- $\pi(n_1) = (move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4)))$
- $\pi(n_2)$

# Continuation of a plan

If $\pi$ is a plan that makes arrival at $n$ possible, $\pi(n)$ **is the continuation of** $\pi$ **at** $n$




Let $\pi$ be the following plan available at $n_0$:

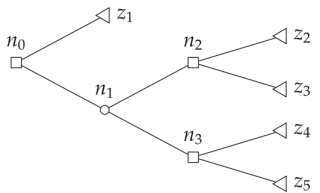
$$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4))))$$

Then:

- $\pi(n_1) = (move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4)))$
- $\pi(n_2) = move(z_2) \wedge stay(z_2)$
- $\pi(n_3)$

# Continuation of a plan

If $\pi$ is a plan that makes arrival at $n$ possible, $\pi(n)$ **is the continuation of** $\pi$ **at** $n$



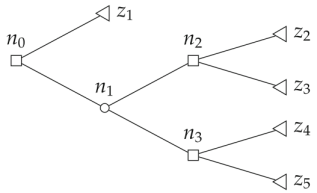

Let $\pi$ be the following plan available at $n_0$:

$$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4))))$$

Then:

- $\pi(n_1) = (move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4)))$
- $\pi(n_2) = move(z_2) \wedge stay(z_2)$
- $\pi(n_3) = move(z_4) \wedge stay(z_4)$
- $\pi(z_i) = stay(z_i)$

# Dynamic consistency




IF the following plan $\pi$ is desirable

$$move(n_1) \wedge ((move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4))))$$

THEN all of the following continuations of $\pi$ are desirable

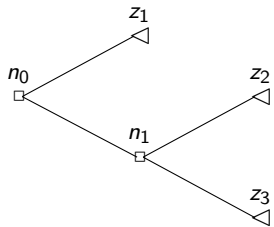
- $\pi(n_1) = (move(n_2) \rightarrow (move(z_2) \wedge stay(z_2))) \wedge (move(n_3) \rightarrow (move(z_4) \wedge stay(z_4)))$
- $\pi(n_2) = move(z_2) \wedge stay(z_2)$
- $\pi(n_3) = move(z_4) \wedge stay(z_4)$
- $\pi(z_i) = stay(z_i)$

# Discussion

- Dynamic consistency tells us that IF $\pi$ is desirable, THEN $\pi(n)$ is desirable. But what about the converse implication?

# Discussion

- Dynamic consistency tells us that IF $\pi$ is desirable, THEN $\pi(n)$ is desirable. But what about the converse implication?
  - Given that a rational agent may have incomplete preferences, it may be that a rational agent comes to consider admissible the continuation of an initially disfavored plan




$z_1$ and $z_3$ are incomparable
$z_2$ and $z_3$ are incomparable
$z_2$ is a tiny bit less preferable than $z_1$

$\pi = \mathit{move}(n_1) \wedge (\mathit{move}(z_2) \wedge \mathit{stay}(z_2))$ is inadmissible
$\pi(n_1) = \mathit{move}(z_2) \wedge \mathit{stay}(z_2)$ is admissible

# Discussion

- Why should we take dynamic consistency to be a rationality constraint?
  - "[I]f our initial endorsement of plan $A$ over plan $B$ is to be genuine, it presumably factors in all the contingencies we are aware of; thus, our endorsement "plan $A$ is better than plan $B$" should not change at our whim." [p.6586]
  - Dynamic inconsistency opens decision makers up to dynamic Dutch books

# Dynamic consistency and desirability maximization

Assume that the agent is a **desirability maximizer** in every decision tree $T$:

For every node $n$ in $T$ and plan $\pi$ available at $n$, $\pi$ is desirable iff,
$value_n(\pi) \geqslant value_n(\pi')$, for all other plans $\pi'$ available at $n$

*Remark.* $value_n(\pi)$ is the desirability value of $\pi$ at node $n$. Desirability values are updated by conditionalization.

# Dynamic consistency and desirability maximization

Assume that the agent is a **desirability maximizer** in every decision tree $T$:

For every node $n$ in $T$ and plan $\pi$ available at $n$, $\pi$ is desirable iff,
$value_n(\pi) \geqslant value_n(\pi')$, for all other plans $\pi'$ available at $n$

*Remark.* $value_n(\pi)$ is the desirability value of $\pi$ at node $n$. Desirability values are updated by conditionalization.

- Are desirability maximizers dynamically consistent?

# Dynamic consistency and desirability maximization

Assume that the agent is a **desirability maximizer** in every decision tree $T$:

For every node $n$ in $T$ and plan $\pi$ available at $n$, $\pi$ is desirable iff,
$value_n(\pi) \geqslant value_n(\pi')$, for all other plans $\pi'$ available at $n$

*Remark.* $value_n(\pi)$ is the desirability value of $\pi$ at node $n$. Desirability values are updated by conditionalization.

- Are desirability maximizers dynamically consistent? Yes, if planning conditionals are construed as indicative conditionals.

# Indicative *vs* subjunctive conditionals

- If Shakespeare did not write Hamlet, someone else did.
- If Shakespeare had not written Hamlet, someone else would have.

It is not clear how the line between the two types of conditionals should be drawn

J. Bennet. *A Philosophical Guide to Conditionals*. Clarendon Press, 2003.

# Planning conditionals as indicative conditionals: preliminaries

1. Indicative conditionals ($\mapsto$) *vs* subjunctive conditionals ($\boxright$):
   - $A \mapsto B$ says that, if we learned that $A$ was actually true, $B$ would be true.
     - We fix the supposition that $A$ is actually true.
   - $A \boxright B$ says that, if $A$ were true, then $B$ would be true.
     - We do not fix the supposition that $A$ is actually true.

# Planning conditionals as indicative conditionals: preliminaries

1. Indicative conditionals ($\mapsto$) *vs* subjunctive conditionals ($\boxright$):
   - $A \mapsto B$ says that, if we learned that $A$ was actually true, $B$ would be true.
     - We fix the supposition that $A$ is actually true.
   - $A \boxright B$ says that, if $A$ were true, then $B$ would be true.
     - We do not fix the supposition that $A$ is actually true.
2. Plans *vs* strategies in extensive form games.
   - A plan specifies a unique move *for every choice node that the agent could reach*, given execution of earlier portions of the plan.
   - A player's strategy in an extensive form game is a *complete* contingency plan: it specifies a unique move *for every choice node of the player*.

# Planning conditionals as indicative conditionals: preliminaries

1. Indicative conditionals ($\mapsto$) *vs* subjunctive conditionals ($\boxright$):
   - $A \mapsto B$ says that, if we learned that $A$ was actually true, $B$ would be true.
     - We fix the supposition that $A$ is actually true.
   - $A \boxright B$ says that, if $A$ were true, then $B$ would be true.
     - We do not fix the supposition that $A$ is actually true.
2. Plans *vs* strategies in extensive form games.
   - A plan specifies a unique move *for every choice node that the agent could reach*, given execution of earlier portions of the plan.
   - A player's strategy in an extensive form game is a *complete* contingency plan: it specifies a unique move *for every choice node of the player*.
3. The argument does not undermine the importance of subjunctive conditionals in the context of rational planning and decision making.

# Planning conditionals as indicative conditionals: preliminaries

*Note that we in no way mean here to deny the importance of subjunctive supposition in the context of rational planning and decision making. Causal decision theory, for example, may well be right to view the practical merits of a plan in terms of its expected desirability under the subjunctive supposition of its implementation. What we deny is simply that the planning conditional itself should be viewed subjunctively.* (fn 16)

# Planning conditionals as indicative conditionals: argument #1

**Example:** Ann is considering the possibility that she will be offered a job at a prestigious law firm and is evaluating the desirability of accepting such an offer, under the supposition that it is made. Ann suffers from terribly low self-esteem and hence is very confident that she will not be offered the position.

# Planning conditionals as indicative conditionals: argument #1

**Example:** Ann is considering the possibility that she will be offered a job at a prestigious law firm and is evaluating the desirability of accepting such an offer, under the supposition that it is made. Ann suffers from terribly low self-esteem and hence is very confident that she will not be offered the position.

- If Ann were, shockingly, to learn that she has been offered the job, the most likely explanation would be that the job was not as serious as she had supposed and so not worth accepting.

# Planning conditionals as indicative conditionals: argument #1

**Example:** Ann is considering the possibility that she will be offered a job at a prestigious law firm and is evaluating the desirability of accepting such an offer, under the supposition that it is made. Ann suffers from terribly low self-esteem and hence is very confident that she will not be offered the position.

- If Ann were, shockingly, to learn that she has been offered the job, the most likely explanation would be that the job was not as serious as she had supposed and so not worth accepting.
- Under these circumstances, Ann may judge accepting the offer as desirable under the subjunctive supposition of its being offered but not under the indicative supposition of its being offered.
  - $\mathit{offer} \mapsto \mathit{not\_accept} \geqslant \mathit{offer} \mapsto \mathit{accept}$
  - $\mathit{offer} \mathrel{\square\!\!\rightarrow} \mathit{accept} \geqslant \mathit{acceptance} \mathrel{\square\!\!\rightarrow} \mathit{not\_go}$

# Planning conditionals as indicative conditionals: argument #1

*This example hopefully brings out why planning conditionals are best understood as indicative rather than subjunctive.*

# Planning conditionals as indicative conditionals: argument #1

*This example hopefully brings out why planning conditionals are best understood as indicative rather than subjunctive. In forming a contingency plan, an agent is considering what to do if, as a matter of fact, various contingencies are found to obtain. When you consider what to do if you are offered the job, you are considering what to do if you in fact learn that you are offered the job.*

# Planning conditionals as indicative conditionals: argument #1

*This example hopefully brings out why planning conditionals are best understood as indicative rather than subjunctive. In forming a contingency plan, an agent is considering what to do if, as a matter of fact, various contingencies are found to obtain. When you consider what to do if you are offered the job, you are considering what to do if you in fact learn that you are offered the job. Strictly counterfactual worlds are of no concern to you, and the counterfactual conditional provides no direct practical guidance.*

# Planning conditionals as indicative conditionals: argument #1

*This example hopefully brings out why planning conditionals are best understood as indicative rather than subjunctive. In forming a contingency plan, an agent is considering what to do if, as a matter of fact, various contingencies are found to obtain. When you consider what to do if you are offered the job, you are considering what to do if you in fact learn that you are offered the job. Strictly counterfactual worlds are of no concern to you, and the counterfactual conditional provides no direct practical guidance. In forming a plan, you are determining how to respond to the different bodies of evidence you might be exposed to, and not how to respond to counterfactual possibilities, which is impossible for an agent located in one world to do. Hence, an indicative reading seems most appropriate given the role planning conditionals are meant to play in the practical deliberation of agents. (pp.6585-99)*

# Some distinctions...

In planning problems, we use conditionals in order to:

- Describe possible outcomes of deliberation:

  *If they offer me the job, I will accept/not accept the offer*

  *If my flight to DC is delayed, I will take the metro/a Uber*

- Evaluate plans:

  *If they offer me the job, the job will not be serious*; so, *if they offer me the job and I accept, I will not have a serious job*

  *If my flight to DC is delayed, there will be only one metro an hour when I arrive*; so, *if my flight to DC is delayed and I take the metro, I will arrive at home very late.*

# Planning conditionals as indicative conditionals: argument #2

*Suppose that plans involve counterfactual contingencies, i.e. they specify moves at nodes that are not reached if the decision maker follows her plan. This can happen, for example, if the agent makes a mistake or acts irrationally at some node. If this happens, though, it is not clear why the agent should be dynamically consistent along the "counterfactual" paths of the decision tree. For then she might learn something about herself that could overturn her initial evaluations of plans.*

# Dynamic consistency of desirability maximization

Assume that indicative conditionals satisfy at least the following properties:

1. **Indicative Property**
   $value_n(\varphi \mapsto \psi) \geqslant value_n(\varphi \mapsto \psi')$ iff $value_n(\varphi \wedge \psi) \geqslant value_n(\varphi \wedge \psi')$
2. **Additivity**
   Where $\{\varphi_i\}$ is a partition, $value_n(\bigwedge_i(\varphi_i \mapsto \psi_i)) = \Sigma_i value_n(\varphi_i \mapsto \psi_i)$

R. Bradley. *Decision theory with a human face*. Cambridge University Press, ch8, 2017.

# Dynamic consistency of desirability maximization

**Lemma 1.** For any decision tree $T$, if $n$ is a **choice node** in $T$, $n$ a node in $T$ that precedes $n'$, and $\pi$ and $\pi'$ are plans available at $n$ consistent with $move(n')$, then

$$value_n(\pi) \geqslant value_n(\pi') \text{ iff } value_{n'}(\pi) \geqslant value_{n'}(\pi')$$

I.e. the relative desirabilities of plans never shift following choice nodes.

**Theorem.** If the planning conditional satisfies the Indicative Property and Additivity, then desirability maximization is dynamically consistent.

# Proof of the main theorem

**Theorem.** If the planning conditional satisfies the Indicative Property and Additivity, then desirability maximization is dynamically consistent.

Take a non-terminal node $n$ in a decision tree $T$ and let $\pi$ be a desirability maximal plan at $n$, i.e., for all plans $\pi'$ available at $n$, $value_n(\pi') \leqslant value_n(\pi)$.

**TBS**:
For all nodes $n'$ s.t. $n$ precedes $n'$, $\pi(n')$ (when defined) is a desirability maximal plan.

**Proof**:
If $n$ is a choice node, $\pi(n')$ is a desirability maximal plan by Lemma 1.

So, let us assume that $n$ is a nature node.

# Proof of the main theorem

$\pi$ is a desirability maximal plan available at a nature node $n$. TBS: for all $n'$ s.t. $n$ precedes $n'$, $\pi(n')$ is a desirability maximal plan.

1. $\pi$ has the form $\bigwedge_i (move(n_i) \mapsto \pi(n_i))$ where $n$ precedes all of the $n_i$

# Proof of the main theorem

$\pi$ is a desirability maximal plan available at a nature node $n$. TBS: for all $n'$ s.t. $n$ precedes $n'$, $\pi(n')$ is a desirability maximal plan.

1. $\pi$ has the form $\bigwedge_i(move(n_i) \mapsto \pi(n_i))$ where $n$ precedes all of the $n_i$
2. Additivity: $value_n(\bigwedge_i(move(n_i) \mapsto \pi(n_i))) = \Sigma_i value_n((move(n_i) \mapsto \pi(n_i)))$

# Proof of the main theorem

$\pi$ is a desirability maximal plan available at a nature node $n$. TBS: for all $n'$ s.t. $n$ precedes $n'$, $\pi(n')$ is a desirability maximal plan.

1. $\pi$ has the form $\bigwedge_i(move(n_i) \mapsto \pi(n_i))$ where $n$ precedes all of the $n_i$
2. Additivity: $value_n(\bigwedge_i(move(n_i) \mapsto \pi(n_i))) = \Sigma_i value_n((move(n_i) \mapsto \pi(n_i)))$
3. Since $\pi$ is a desirability maximal plan, for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \mapsto \pi(n_i)) \geqslant value_n(move(n_i) \mapsto \pi'(n_i))$

# Proof of the main theorem

$\pi$ is a desirability maximal plan available at a nature node $n$. TBS: for all $n'$ s.t. $n$ precedes $n'$, $\pi(n')$ is a desirability maximal plan.

1. $\pi$ has the form $\bigwedge_i(move(n_i) \mapsto \pi(n_i))$ where $n$ precedes all of the $n_i$
2. Additivity: $value_n(\bigwedge_i(move(n_i) \mapsto \pi(n_i))) = \Sigma_i value_n((move(n_i) \mapsto \pi(n_i)))$
3. Since $\pi$ is a desirability maximal plan, for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \mapsto \pi(n_i)) \geqslant value_n(move(n_i) \mapsto \pi'(n_i))$
4. Indicative Property: for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \wedge \pi(n_i)) \geqslant value_n(move(n_i) \wedge \pi'(n_i))$

# Proof of the main theorem

$\pi$ is a desirability maximal plan available at a nature node $n$. TBS: for all $n'$ s.t. $n$ precedes $n'$, $\pi(n')$ is a desirability maximal plan.

1. $\pi$ has the form $\bigwedge_i(move(n_i) \mapsto \pi(n_i))$ where $n$ precedes all of the $n_i$
2. Additivity: $value_n(\bigwedge_i(move(n_i) \mapsto \pi(n_i))) = \Sigma_i value_n((move(n_i) \mapsto \pi(n_i)))$
3. Since $\pi$ is a desirability maximal plan, for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \mapsto \pi(n_i)) \geqslant value_n(move(n_i) \mapsto \pi'(n_i))$
4. Indicative Property: for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \wedge \pi(n_i)) \geqslant value_n(move(n_i) \wedge \pi'(n_i))$
5. Hence: $value_n(move(n_i) \wedge \pi(n_i)) - value_n(move(n_i)) \geqslant value_n(move(n_i) \wedge \pi'(n_i)) - value_n(move(n_i))$

# Proof of the main theorem

$\pi$ is a desirability maximal plan available at a nature node $n$. TBS: for all $n'$ s.t. $n$ precedes $n'$, $\pi(n')$ is a desirability maximal plan.

1. $\pi$ has the form $\bigwedge_i(move(n_i) \mapsto \pi(n_i))$ where $n$ precedes all of the $n_i$
2. Additivity: $value_n(\bigwedge_i(move(n_i) \mapsto \pi(n_i))) = \Sigma_i value_n((move(n_i) \mapsto \pi(n_i)))$
3. Since $\pi$ is a desirability maximal plan, for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \mapsto \pi(n_i)) \geqslant value_n(move(n_i) \mapsto \pi'(n_i))$
4. Indicative Property: for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \wedge \pi(n_i)) \geqslant value_n(move(n_i) \wedge \pi'(n_i))$
5. Hence: $value_n(move(n_i) \wedge \pi(n_i)) - value_n(move(n_i)) \geqslant value_n(move(n_i) \wedge \pi'(n_i)) - value_n(move(n_i))$
6. Conditional desirability: $value_n(\pi(n_i) \mid move(n_i)) \geqslant value_n(\pi'(n_i) \mid move(n_i))$

# Proof of the main theorem

$\pi$ is a desirability maximal plan available at a nature node $n$. TBS: for all $n'$ s.t. $n$ precedes $n'$, $\pi(n')$ is a desirability maximal plan.

1. $\pi$ has the form $\bigwedge_i(move(n_i) \mapsto \pi(n_i))$ where $n$ precedes all of the $n_i$
2. Additivity: $value_n(\bigwedge_i(move(n_i) \mapsto \pi(n_i))) = \Sigma_i value_n((move(n_i) \mapsto \pi(n_i)))$
3. Since $\pi$ is a desirability maximal plan, for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \mapsto \pi(n_i)) \geqslant value_n(move(n_i) \mapsto \pi'(n_i))$
4. Indicative Property: for all $n_i$ and for all $\pi'$ available at $n$, $value_n(move(n_i) \wedge \pi(n_i)) \geqslant value_n(move(n_i) \wedge \pi'(n_i))$
5. Hence: $value_n(move(n_i) \wedge \pi(n_i)) - value_n(move(n_i)) \geqslant value_n(move(n_i) \wedge \pi'(n_i)) - value_n(move(n_i))$
6. Conditional desirability: $value_{n_i}(\pi(n_i)) \geqslant value_{n_i}(\pi'(n_i))$

# Take home ideas

- With respect to static decision problems, dynamic decision problems may involve additional rationality criteria (like dynamic consistency). It is not obvious that these criteria interact in an appropriate way with the rationality criteria for static decision problems (like desirability maximization).

# Take home ideas

- With respect to static decision problems, dynamic decision problems may involve additional rationality criteria (like dynamic consistency). It is not obvious that these criteria interact in an appropriate way with the rationality criteria for static decision problems (like desirability maximization).

- When we decide which plan to perform, we use conditionals for:

  1. describing the plans that are possible (the possible outcomes of deliberation):
     *I will do A and then, if X happens, I will do B and, if Y happens, I will do C*
  2. assessing the plans in question:
     *If X happens and I do B, then it will be the case that X'*
     *If X happens and I do B', then it will be the case that X"*
     Since $X'$ is better than $X''$, if $X$ happens, I should do $B$.

# Choice driven counterfactuals in branching time

IC & Eric Pacuit. *Choice driven counterfactuals*. JPL, 51, pp. 297–345, 2022.

# Choice-driven counterfactuals

Suppose that the charge nurse puts the wrong medications in Bob's pill organizer and that the intern gives them to Bob, who has an allergic reaction.

# Choice-driven counterfactuals

Suppose that the charge nurse puts the wrong medications in Bob's pill organizer and that the intern gives them to Bob, who has an allergic reaction. Who is responsible for Bob's allergic reaction?

# Choice-driven counterfactuals

Suppose that the charge nurse puts the wrong medications in Bob's pill organizer and that the intern gives them to Bob, who has an allergic reaction. Who is responsible for Bob's allergic reaction?

- Who did causally contribute to the allergic reaction?

# Choice-driven counterfactuals

Suppose that the charge nurse puts the wrong medications in Bob's pill organizer and that the intern gives them to Bob, who has an allergic reaction. Who is responsible for Bob's allergic reaction?

- Who did causally contribute to the allergic reaction?
- Would the intern have given the wrong medications to Bob had the charge nurse put the right medications in his pill organizer?

# Choice-driven counterfactuals

Suppose that the charge nurse puts the wrong medications in Bob's pill organizer and that the intern gives them to Bob, who has an allergic reaction. Who is responsible for Bob's allergic reaction?

- Who did causally contribute to the allergic reaction?
- Would the intern have given the wrong medications to Bob had the charge nurse put the right medications in his pill organizer?

The latter question involves a **choice-driven counterfactual**, i.e., a counterfactual whose semantic value depends on the choices that the agents are expected to make (on the agents' *default choice behavior*).

Choice-driven counterfactuals are important for, e.g., determining responsibility, making plans for the future, strategic reasoning about how our choices influence the choices of others.

# More examples

- Ann would have picked up the kids if her husband hadn't.
- If David had bet tails, Max wouldn't have kept playing.
- If Alice hadn't screamed, the thief wouldn't have shot her.
- If the charge nurse hadn't put the wrong medications on the desk, the intern wouldn't have given them to the patient.

# Our aim

We study the semantics and logical properties of choice-driven counterfactuals. To do this, we merge **STIT logic** (the logic of **S**eeing **T**o **I**t **T**hat) with the logic of counterfactuals due to Stalnanker (1968) and Lewis (1973).

# A bit of context...

## STIT logic:

N. Belnap, M. Perloff, M. Xu. *Facing the Future*. OUP, 2001.

J. Horty. *Agency and Deontic Logic*. OUP, 2001.

## Counterfactuals in STIT:

M. Xu. *Causation in Branching Time (I): Transitions, Events and Causes*. Synthese, 112(2): 137 - 192..

J. Horty. *Agency and Deontic Logic (Chapter 4)*. OUP, 2001.

## Counterfactuals in branching time:

R. Thomason & A. Gupta. *A Theory of Conditionals in the Context of Branching Time*. The Philosophical Review, 89(1), pp. 65-90, 1980.

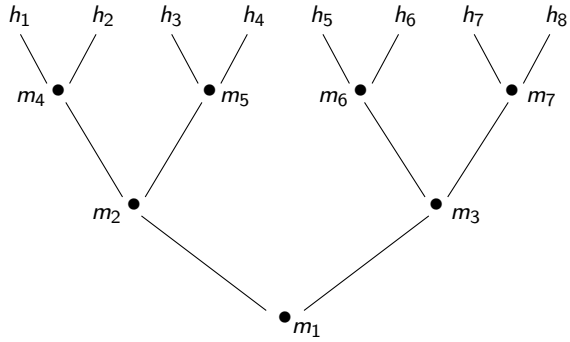
T. Placek & T. Müller. *Counterfactuals and historical possibility*. Synthese, 154(2), pp. 173-197.

# STIT semantics

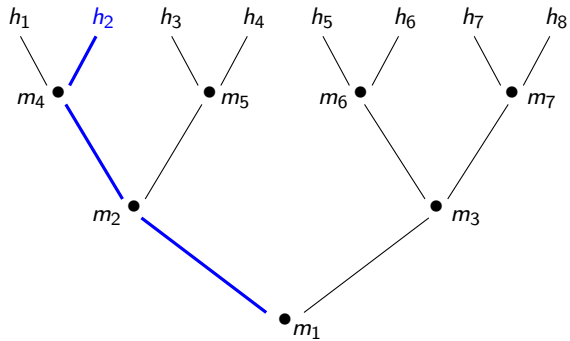
STIT is a formal theory of **agency** cast against the background of a theory of **indeterministic time**.

- STIT models consist of two components:
  - A branching time structure
  - Agents' choices

# Discrete branching time structures

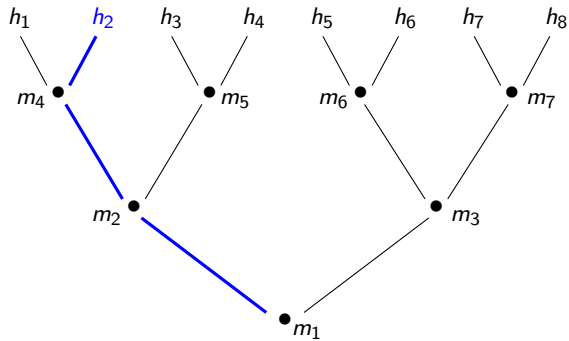


# Discrete branching time structures




A **history** $h$ is a maximally linearly ordered set of moments.

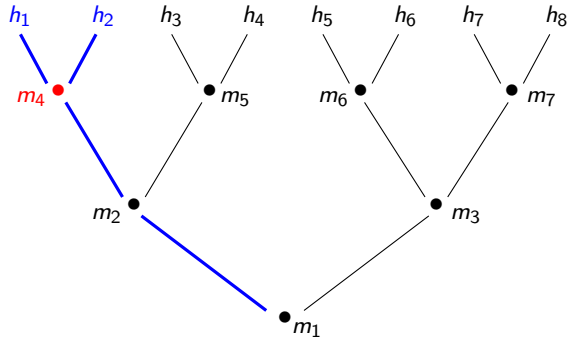
# Discrete branching time structures




A **history** $h$ is a maximally linearly ordered set of moments.

$H_m = \{h \mid m \in h\}$ is the set of histories passing through $m$.

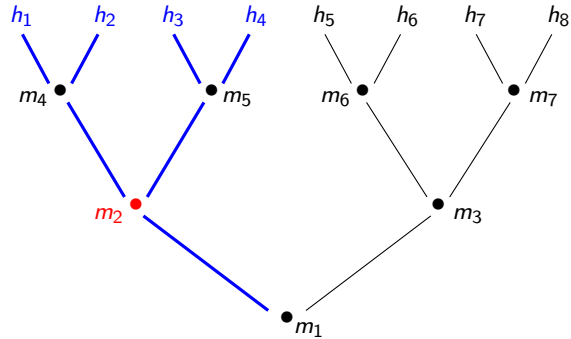
# Discrete branching time structures




A **history** $h$ is a maximally linearly ordered set of moments.

$H_m = \{h \mid m \in h\}$ is the set of histories passing through $m$.

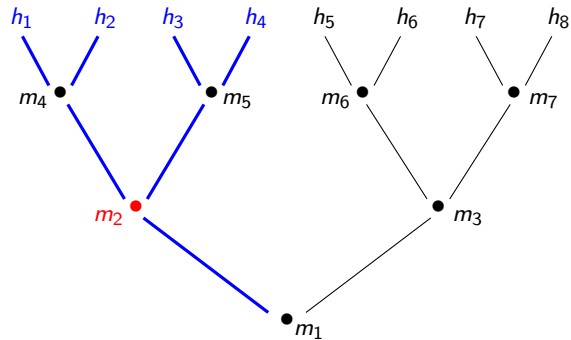
# Discrete branching time structures




A **history** $h$ is a maximally linearly ordered set of moments.

$H_m = \{h \mid m \in h\}$ is the set of histories passing through $m$.

# Discrete branching time structures

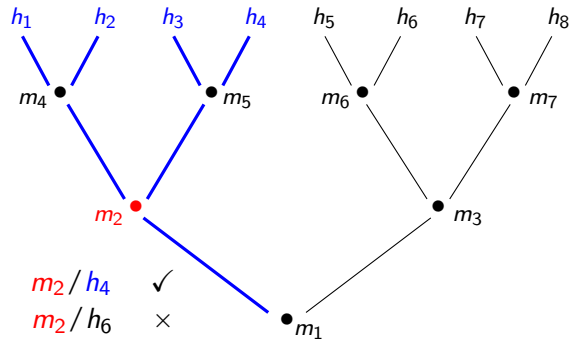



A **history** $h$ is a maximally linearly ordered set of moments.

$H_m = \{h \mid m \in h\}$ is the set of histories passing through $m$.

An **index** is a pair $m/h$ where $h \in H_m$.

# Discrete branching time structures

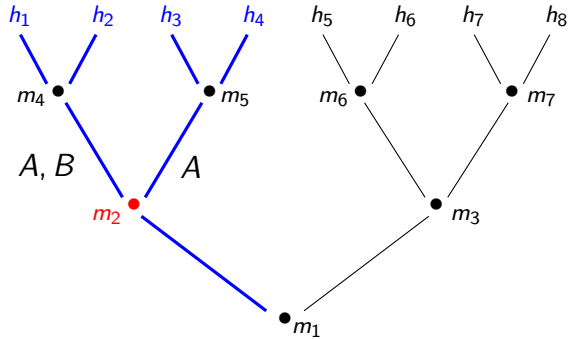



A **history** $h$ is a maximally linearly ordered set of moments.

$H_m = \{h \mid m \in h\}$ is the set of histories passing through $m$.

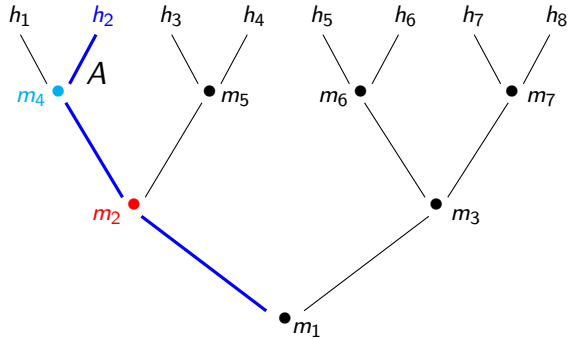
An **index** is a pair $m/h$ where $h \in H_m$.

# Historic necessity, "next", and yesterday operators




$\mathcal{M}, m/h \models \Box A$ iff, for all $h' \in H_m$, $\mathcal{M}, m/h' \models A$

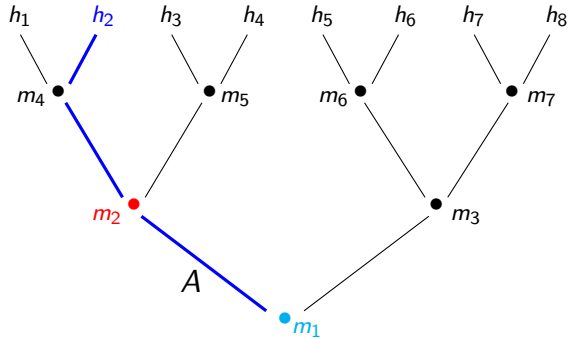
# Historic necessity, "next", and yesterday operators




$\mathcal{M}, m/h \models \Box A$ iff, for all $h' \in H_m$, $\mathcal{M}, m/h' \models A$

$\mathcal{M}, m/h \models \mathsf{X}A$ iff $\mathcal{M}, succ_h(m)/h \models A$

# Historic necessity, "next", and yesterday operators

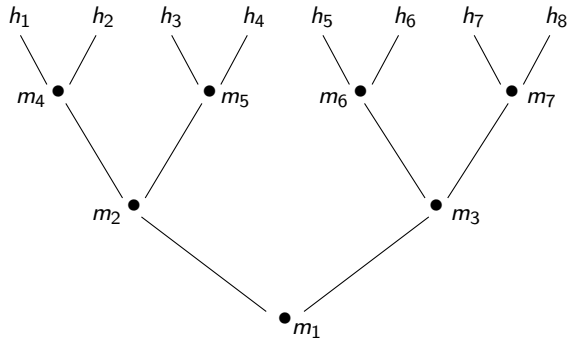



$\mathcal{M}, m/h \models \Box A$ iff, for all $h' \in H_m$, $\mathcal{M}, m/h' \models A$

$\mathcal{M}, m/h \models \mathsf{X}A$ iff $\mathcal{M}, succ_h(m)/h \models A$

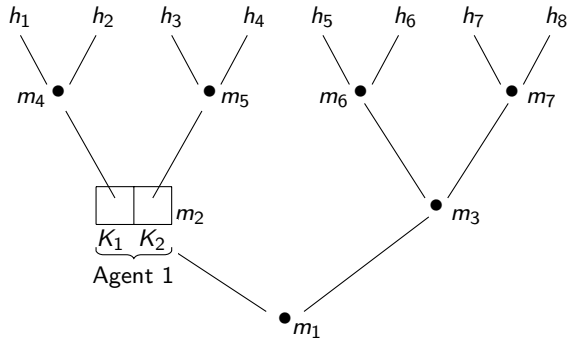
$\mathcal{M}, m/h \models \mathsf{Y}A$ iff $m = m_1$ or $\mathcal{M}, pred(m)/h \models A$

# Choices and STIT operators



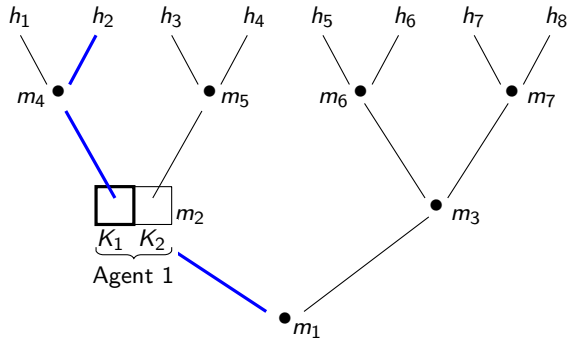

Choices available to an agent $i$ at $m$ are a partition of $H_m$

# Choices and STIT operators



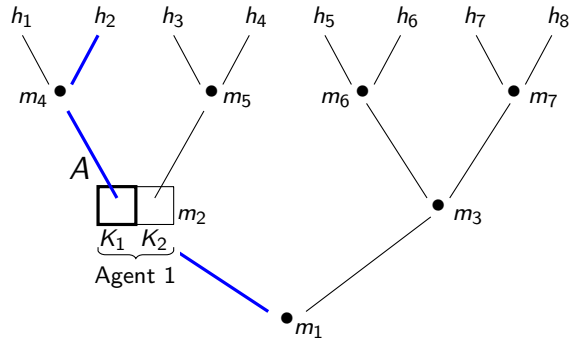

Choices available to an agent $i$ at $m$ are a partition of $H_m$

# Choices and STIT operators




Choices available to an agent $i$ at $m$ are a partition of $H_m$

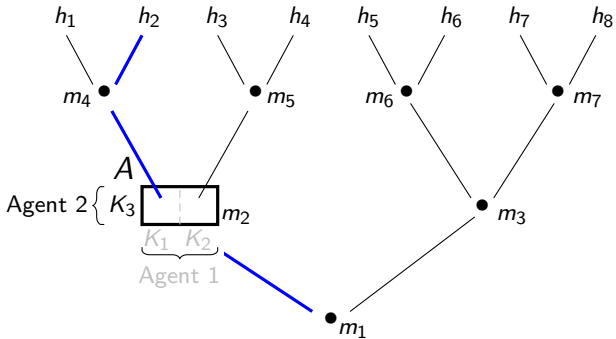
# Choices and STIT operators




Choices available to an agent $i$ at $m$ are a partition of $H_m$

$\mathcal{M}, m/h \models [i \ stit\ ]A$ iff, for all $h' \in [h]_i^m$, $\mathcal{M}, m/h' \models A$

# Choices and STIT operators




Choices available to an agent $i$ at $m$ are a partition of $H_m$

$\mathcal{M}, m/h \models [i \ stit\ ]A$ iff, for all $h' \in [h]_i^m$, $\mathcal{M}, m/h' \models A$

# Why extending STIT

1. David decides whether to play with Max or Maxine.
2. He bets heads or tails.
3. The person nominated by David flips a coin.
4. David wins iff his bet matches the outcome of the coin flip.
5. Max wins iff David loses.
6. Maxine always wins.

# Why extending STIT

1. David decides whether to play with Max or Maxine.
2. He bets heads or tails.
3. The person nominated by David flips a coin.
4. David wins iff his bet matches the outcome of the coin flip.
5. Max wins iff David loses.
6. Maxine always wins.

Both Max and Maxine have two coins, one with heads on each side (H-coin) and one with tails on each side (T-coin). If **Max** has a chance to play, he will choose so as to make David lose. If **Maxine** has a chance to play, she picks one of the coins to flip at random.

# Why extending STIT

1. David decides whether to play with Max or Maxine.
2. He bets heads or tails.
3. The person nominated by David flips a coin.
4. David wins iff his bet matches the outcome of the coin flip.
5. Max wins iff David loses.
6. Maxine always wins.

Both Max and Maxine have two coins, one with heads on each side (H-coin) and one with tails on each side (T-coin). If **Max** has a chance to play, he will choose so as to make David lose. If **Maxine** has a chance to play, she picks one of the coins to flip at random.

**Facts:** after nominating Max, David bets heads and Max flips the T-coin.

# Why extending STIT

The following counterfactual is intuitively true

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

# Why extending STIT

The following counterfactual is intuitively true

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

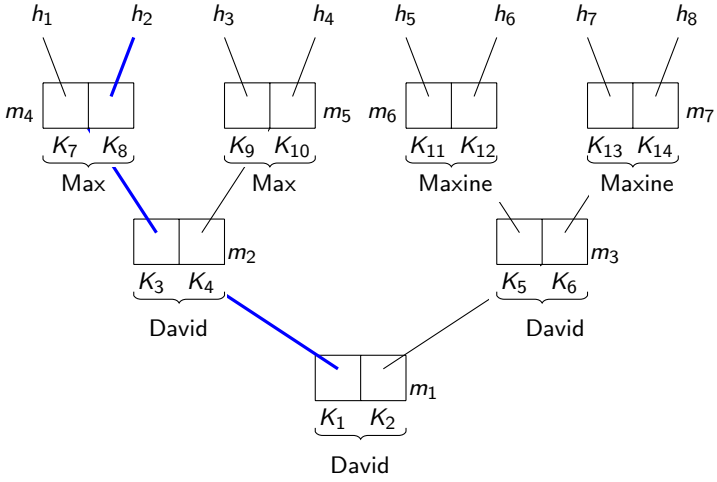
Why? If David had bet tails instead of heads, Max would have flipped the H-coin, thus making David lose.

# Why extending STIT

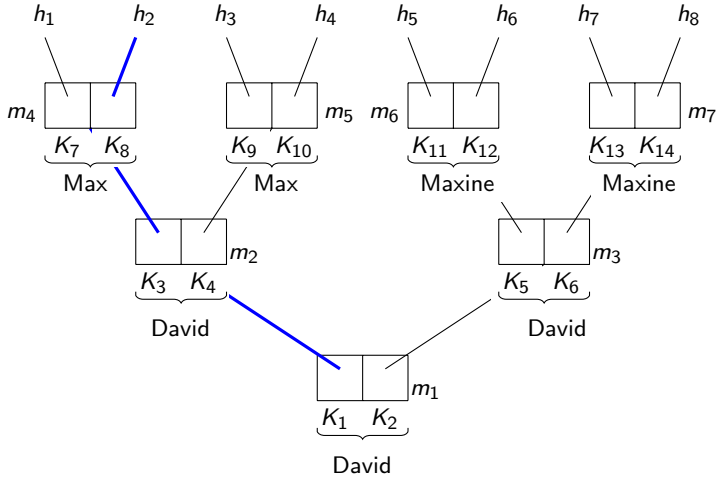
We need a semantics that can represent the following elements:

1. The different ways in which **things could have gone**
   (David could have bet differently)
2. The particular **time** at which an agent makes a choice
   (We consider alternatives where David has *just* bet tails)
3. The **types of action** performed by the agents
   (We consider alternatives where David performs the action type "betting tails")
4. The **default choice behavior** of the agents
   (When we evaluate $C1$ we rely on default assumptions about what Max would have done had David acted differently)

# Extending STIT

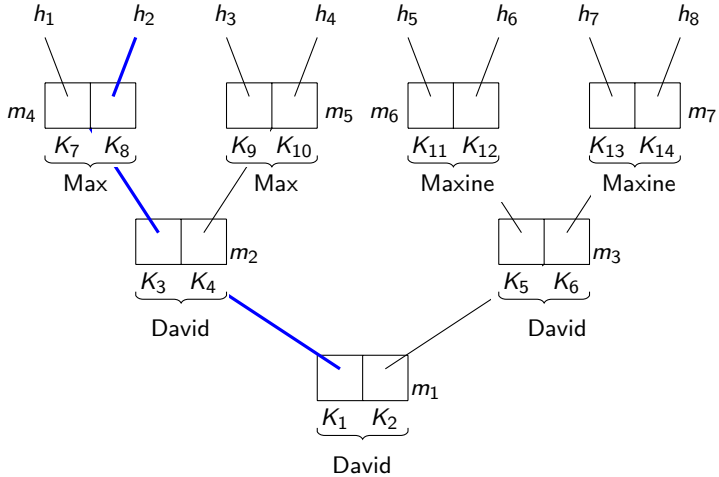


# Extending STIT



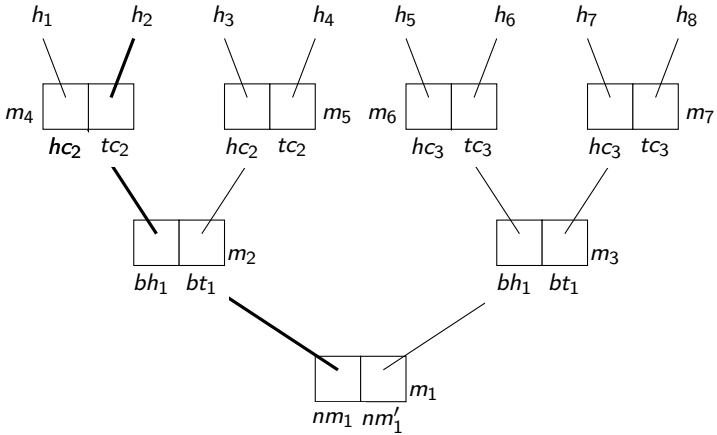

At $m_2 / h_2$, if David had bet tails, then he would have lost.

# Extending STIT



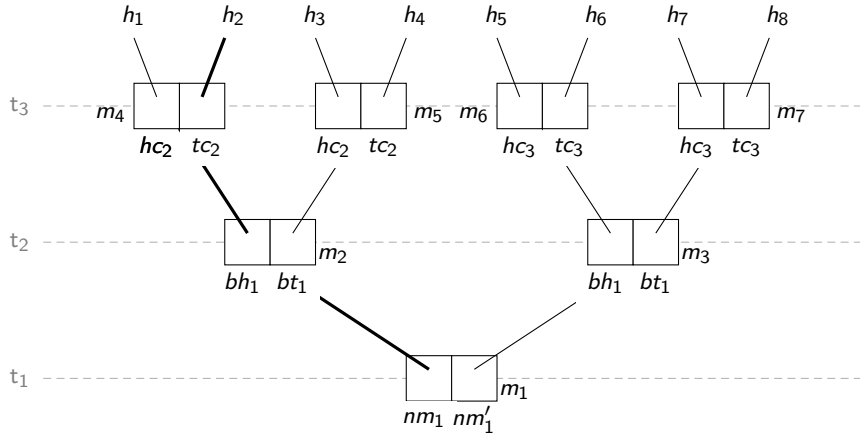

At $m_2/h_2$, if David had bet tails, then he would have lost.

# Extending STIT



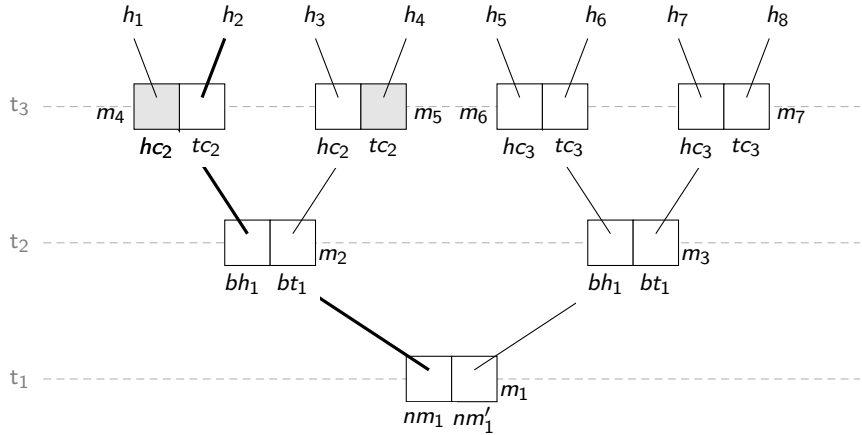

**Step 1:** we label actions with their **types**.

# Extending STIT



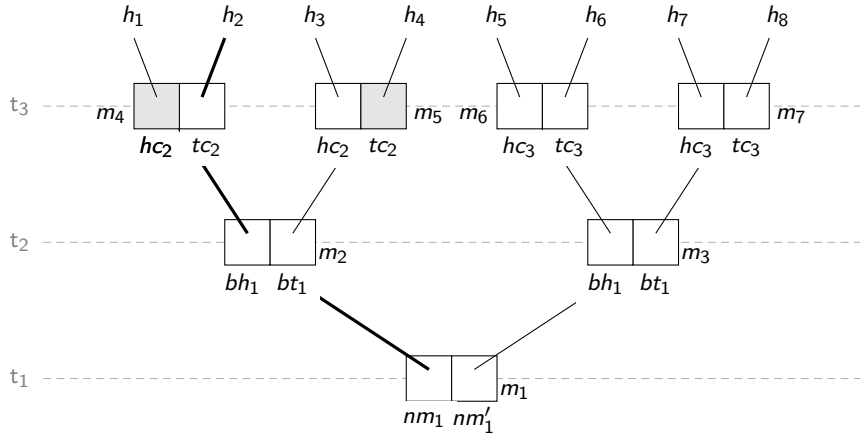

**Step 2:** we "add" **instants** (moments occurring at the same time).

# Extending STIT




**Step 3:** we add **deviant choices** to represent *choice rules*.

# Extending STIT




- $\mathcal{M}, m/h \models do(a_i)$ iff $\mathbf{act}(m/h)(i) = a_i$
- $\mathcal{M}, m/h \models dev(a_i)$ iff $a_i \in \mathbf{dev}(m)$

# Adding counterfactuals

- We introduce formulas of the form $\varphi \mathbin{\Box\!\!\rightarrow} \psi$ (read: if $\varphi$, $\psi$ would be true)

# Adding counterfactuals

- We introduce formulas of the form $\varphi \mathbin{\Box\!\!\rightarrow} \psi$ (read: if $\varphi$, $\psi$ would be true)
- **Aim**: interpreting $\varphi \mathbin{\Box\!\!\rightarrow} \psi$ on our models

# Adding counterfactuals

- We introduce formulas of the form $\varphi \boxrightarrow \psi$ (read: if $\varphi$, $\psi$ would be true)
- **Aim**: interpreting $\varphi \boxrightarrow \psi$ on our models
- **Starting point**: Stalnaker-Lewis semantics

$\varphi \boxrightarrow \psi$ is true at a world $w$ just in case

*either* there is no $\varphi$-world accessible from $w$ (vacuous case),

*or* some $\varphi \wedge \psi$-world accessible from $w$ that is **more similar** to $w$ than any $\varphi \wedge \neg\psi$-world.

# Adding counterfactuals

- We introduce formulas of the form $\varphi \boxrightarrow \psi$ (read: if $\varphi$, $\psi$ would be true)
- **Aim**: interpreting $\varphi \boxrightarrow \psi$ on our models
- **Starting point**: Stalnaker-Lewis semantics

$\varphi \boxrightarrow \psi$ is true at a world $w$ just in case

*either* there is no $\varphi$-world accessible from $w$ (vacuous case),

*or* some $\varphi \wedge \psi$-world accessible from $w$ that is **more similar** to $w$ than any $\varphi \wedge \neg\psi$-world.

- **Question**: relative similarity between indices or histories?

# Relative similarity over histories

Let $\preceq$: $Hist \rightarrow 2^{Hist \times Hist}$ be a function assigning to every history $h$ a relation $\preceq_h$ where $h_1 \preceq_h h_2$ means "$h_1$ is at least as similar to $h$ as $h_2$"

# Truth conditions for $\Box\!\!\rightarrow$

**Stalnaker-Lewis semantics**

$\varphi \Box\!\!\rightarrow \psi$ is true at world $w$ just in case

*either* there is no $\varphi$-world accessible from $w$ (vacuous case),

*or* some $\varphi \wedge \psi$-world accessible from $w$ that is **more similar** to $w$ than any $\varphi \wedge \neg\psi$-world.

# Truth conditions for $\Box\!\!\rightarrow$

**Stalnaker-Lewis semantics adapted**

$\varphi \Box\!\!\rightarrow \psi$ is true at index $m/h$ just in case

*either* there is no $h' \in \mathit{Hist}$ s.t. $t_m/h' \models \varphi$,

*or* there is $h' \in \mathit{Hist}$ s.t. $t_m/h' \models \varphi \wedge \psi$ and,
if $t_m/h'' \models \varphi \wedge \neg\psi$, then $h'' \not\leq_h h'$

$t_m/h'$ is the index consisting of the moment on $h'$ occurring at the time of $m$

# Truth conditions for $\boxrightarrow$

**Stalnaker-Lewis semantics adapted**

$\varphi \boxrightarrow \psi$ is true at index $m/h$ just in case

*either* there is no $h' \in \mathit{Hist}$ s.t. $t_m/h' \models \varphi$,

*or* there is $h' \in \mathit{Hist}$ s.t. $t_m/h' \models \varphi \wedge \psi$ and, if $t_m/h'' \models \varphi \wedge \neg\psi$, then $h'' \not\leq_h h'$

**Hidden assumptions**:

# Truth conditions for $\Box\!\!\rightarrow$

**Stalnaker-Lewis semantics adapted**
$\varphi \Box\!\!\rightarrow \psi$ is true at index $m/h$ just in case

*either* there is no $h' \in \mathit{Hist}$ s.t. $t_m/h' \models \varphi$,

*or* there is $h' \in \mathit{Hist}$ s.t. $t_m/h' \models \varphi \wedge \psi$ and,
if $t_m/h'' \models \varphi \wedge \neg\psi$, then $h'' \not\leq_h h'$

**Hidden assumptions**:

1. the truth value of $\varphi$ and $\psi$ at indices not occurring at the time of evaluation ($t_m$) does not affect the truth-value of $\varphi \Box\!\!\rightarrow \psi$.

# Truth conditions for $\Box\!\!\rightarrow$

**Stalnaker-Lewis semantics adapted**
$\varphi \Box\!\!\rightarrow \psi$ is true at index $m/h$ just in case

*either* there is no $h' \in \mathit{Hist}$ s.t. $t_m/h' \models \varphi$,

*or* there is $h' \in \mathit{Hist}$ s.t. $t_m/h' \models \varphi \wedge \psi$ and,
if $t_m/h'' \models \varphi \wedge \neg\psi$, then $h'' \not\leq_h h'$

**Hidden assumptions**:

1. the truth value of $\varphi$ and $\psi$ at indices not occurring at the time of evaluation ($t_m$) does not affect the truth-value of $\varphi \Box\!\!\rightarrow \psi$.
2. the time of evaluation does not affect the relation of relative similarity between histories.

# Defining $\preceq_h$: Analysis 0

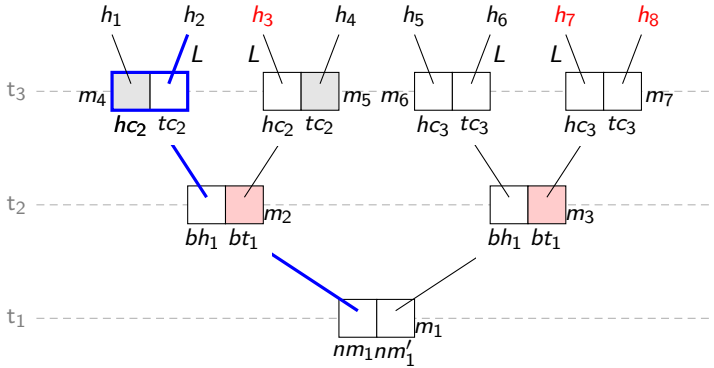
*It is of the first importance to avoid*
*big, widespread, diverse violations of law.*

D. Lewis. *Counterfactual dependence and time's arrow*. Nous, 13(4): 455-476.

- $h_1$ is more similar to $h$ than $h_2$ if fewer deviations from the agents' default choice behavior occur on $h_1$ than on $h_2$:

$$h_1 \prec_h h_2 \text{ iff } n\_dev(h_1) < n\_dev(h_2)$$

# Excluding Analysis 0




We want $m_4/h_2 \models \mathrm{Y}(do(bt_1)) \mathrel{\square\!\!\rightarrow} L$ ("if David had bet tails, he would have lost").

But the histories with the fewest number of deviations on which $\mathrm{Y}(do(bt_1))$ is true at $t_{m_4} = t_3$ are $h_3, h_7, h_8$ and $L$ is false on $h_1$ and $h_8$ at $t_2$...

# Defining $\preceq_h$: Analysis 1

The greater past overlap between $h_2$ and $h_3$ is more important than the equal number of deviations on $h_7$ and $h_8$.

# Defining $\preceq_h$: Analysis 1

The greater past overlap between $h_2$ and $h_3$ is more important than the equal number of deviations on $h_7$ and $h_8$.

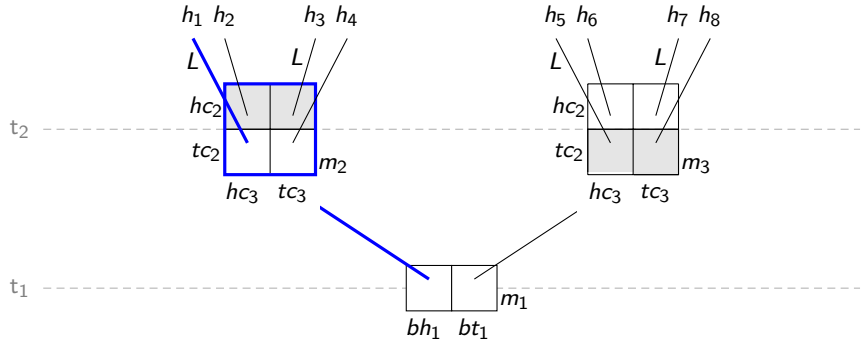
- $h_1 \prec_h h_2$ iff

  **either** $past\_ov(h, h_1) \supset past\_ov(h, h_2)$

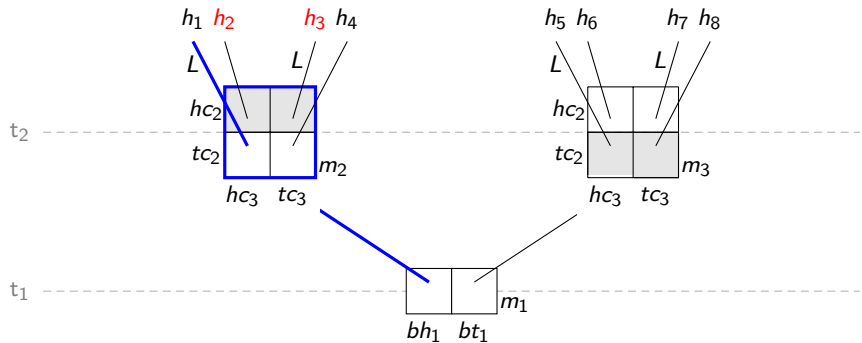
  **or** $past\_ov(h, h_1) = past\_ov(h, h_2)$ and $n\_dev(h_1) < n\_dev(h_2)$

where $past\_ov(h, h') = h \cap h'$

# Excluding Analysis 1



# Excluding Analysis 1




We want $m_2/h_1 \models Y(do(hc_2)) \boxrightarrow \neg L$ ("If Max had flipped the H-coin, David would have won").

But $h_2$ and $h_3$ are the most similar histories to $h_1$ on which Max flips the H-coin at $t_{m_2} = t_2$ and $L$ is true on $h_3$ at $t_2$...

# Proposal 1: Rewind similarity function

The smaller change making $h_2$ branch off from $h_1$ is more important than the equal past overlap between on $h_1$ and $h_2$ and between $h_1$ and $h_3$.

# Proposal 1: Rewind similarity function

The smaller change making $h_2$ branch off from $h_1$ is more important than the equal past overlap between on $h_1$ and $h_2$ and between $h_1$ and $h_3$.

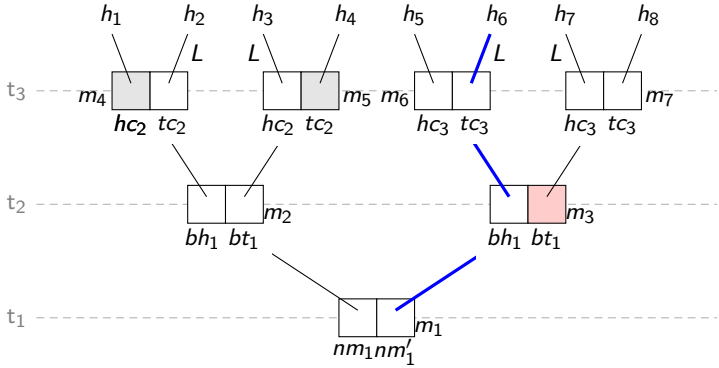
- Define $<^R$: $Hist \to \wp(Hist \times Hist)$ as follows:

  $h_1 <^R_h h_2$ iff either one of the following obtains:

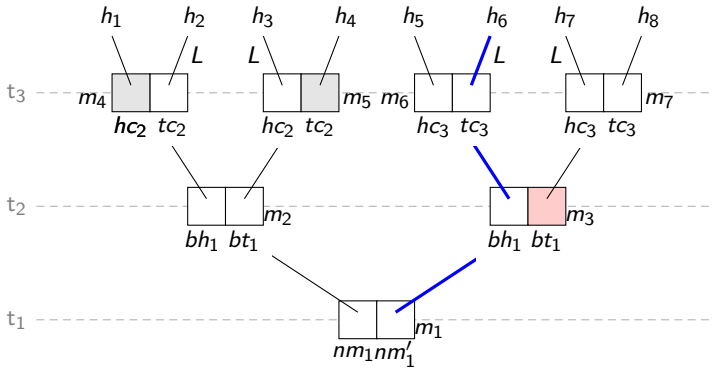
  - $past\_ov(h, h_1) \supset past\_ov(h, h_2)$
  - $past\_ov(h, h_1) = past\_ov(h, h_2)$ and $n\_sep(h, h_1) < n\_sep(h, h_2)$
  - $past\_ov(h, h_1) = past\_ov(h, h_2)$ and $n\_sep(h, h_1) = n\_sep(h, h_2)$ and $n\_dev(h_1) < n\_dev(h_2)$

where $n\_sep(h, h_1)$ counts the number of actions making $h_1$ branch off from $h$.

# Rewind *vs* independence



# Rewind *vs* independence




Would David have won had he bet tails?

# Rewind *vs* independence

**Rewind History**: When we suppose that David bet differently, we *rewind* the course of events to the moment when David bets ($m_1$), intervene on his choice, and then let the future unfold according to the agents' default choice behavior. Since there is no constraint on the coin that Maxine will flip, we only conclude that David **might** win.

D. Lewis.. *Counterfactual dependence and time's arrow*. Nous, 13(4), pp. 455-476, 1979.

# Rewind *vs* independence

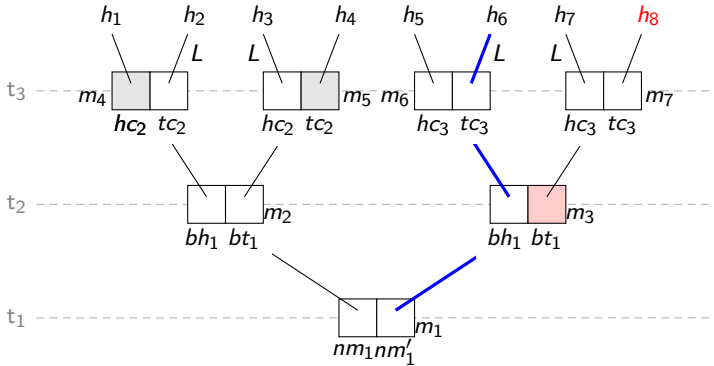
**Rewind History**: When we suppose that David bet differently, we *rewind* the course of events to the moment when David bets ($m_1$), intervene on his choice, and then let the future unfold according to the agents' default choice behavior. Since there is no constraint on the coin that Maxine will flip, we only conclude that David **might** win.

D. Lewis.. *Counterfactual dependence and time's arrow*. Nous, 13(4), pp. 455-476, 1979.

**Assume Causal 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. Since there is no choice rule according to which Maxine's choice depends on David's bet, we conclude that, if he had bet differently, then David **would** have won.

M.A. Slote. *Time in counterfactuals*. The Phil Review, 87(1), pp. 3-27, 1978.

# Proposal 2: idea




The fact that more unconstrained agents act in the same way on $h_6$ and $h_8$ than on $h_6$ and $h_7$ is more important than the equal number of deviations on $h_7$ and $h_8$.

# Proposal 2: Independence similarity functions

Define $\prec^I$: $Hist \rightarrow \wp(Hist \times Hist)$ as follows:

$h_1 \prec^I_h h_2$ iff either one of the following obtains:

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

$n\_unc(h, h_1)$ counts the number of unconstrained agents acting in the same way on $h$ and $h_1$.

# What if deviant actions were performed in the past?

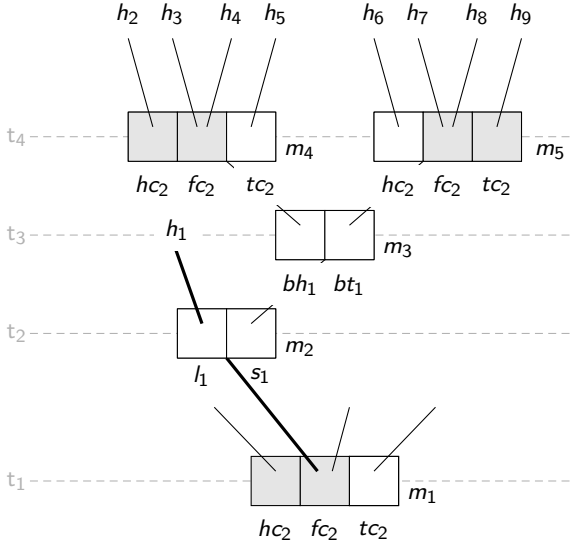
A crucial assumption that both definitions of similarity rely on is that the evaluation of choice-driven counterfactuals depends on the default choice behavior of the agents. Do these definitions still make sense when we evaluate a choice-driven counterfactual on a history where one or more agents behaved deviantly in the past?

# A variation of our example

**Example 4.** Everything is as before, 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.

Is the following counterfactual true?

- If David were to bet heads again, he would lose $(\mathrm{X}do(bh_1) \mathbin{\square\!\!\rightarrow} \mathrm{XX}L)$




- According to our definitions, $(\mathrm{X}do(bh_1) \square\!\!\rightarrow \mathrm{XX}L)$ is true at $m_2/h_1$

# Counterfactual reasoning with past deviations

Given that Max acted deviantly, it is not clear that the previous judgment is correct. In fact, we can reason about what Max would do in the second game in different ways:

1. forget that Max's actual choice was deviant and assume that he is still constrained by his choice rule;
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;
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.

# Connection with backward and forward induction

- **Backward induction**: players ignore past behavior and reason only about their opponents' future moves

# Counterfactual reasoning with past deviations

Given that Max acted deviantly, it is not clear that the previous judgment is correct. In fact, we can reason about what Max would do in the second game in different ways:

1. forget that Max's actual choice was deviant and assume that he is still constrained by his choice rule;
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;
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.

# Connection with backward and forward induction

- **Backward induction**: players ignore past behavior and reason only about their opponents' future moves
- **Forward induction**: players rationalize past behavior and use it as a basis for forming beliefs about future moves

# Counterfactual reasoning with past deviations

Given that Max acted deviantly, it is not clear that the previous judgment is correct. In fact, we can reason about what Max would do in the second game in different ways:

1. forget that Max's actual choice was deviant and assume that he is still constrained by his choice rule;
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;
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.

# What would Max do in the second game?

Given that Max acted deviantly, it is not clear that the previous judgment is correct. In fact, we can reason about what Max would do in the second game in different ways:

1. forget that Max's actual choice was deviant and assume that he is still constrained by his choice rule;
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;
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.

We propose a refinement of our definitions that captures the idea that past deviations can influence the agents' future choices (options 2 or 3).

# Thank you!!

# Course material and contacts

Course website:

https://pacuit.org/esslli2022/conditionals-games/

Our email addresses:

epacuit@umd.edu
icanavot@umd.edu

Our websites:

https://pacuit.org/
https://sites.google.com/view/ilariacanavotto/