Counterfactuals in Minds and Machines

Tobias Gerstenberg

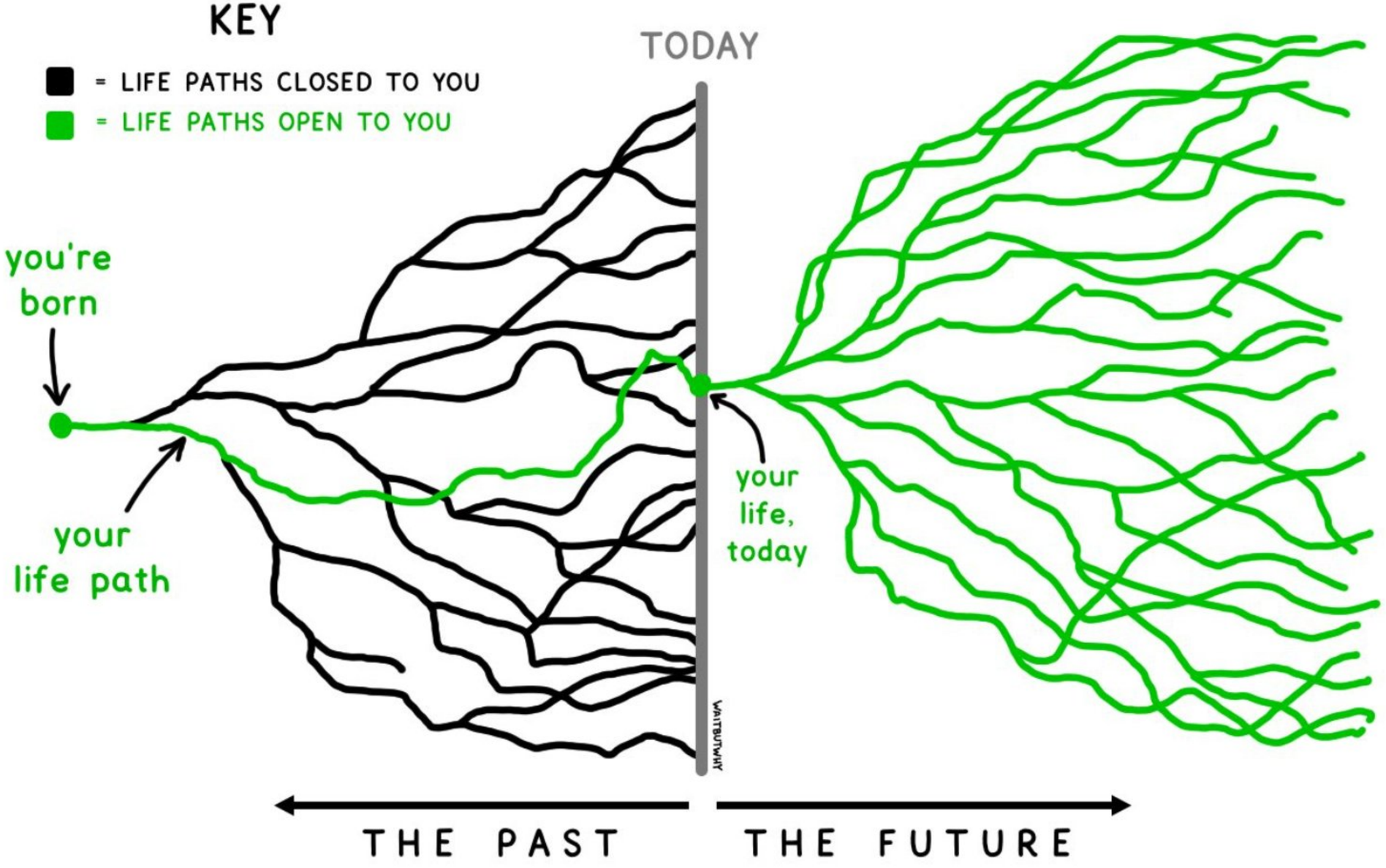
Manuel Gomez-Rodriguez

Stratis Tsirtsis

Stanford University

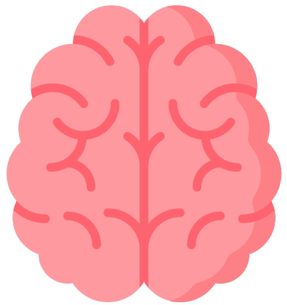
MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

UAI, Rio de Janeiro
21/7/2025


KEY
= LIFE PATHS CLOSED TO YOU
= LIFE PATHS OPEN TO YOU
TODAY
you're born
your life path
your life, today
WAITBUTWHY
THE PAST
THE FUTURE

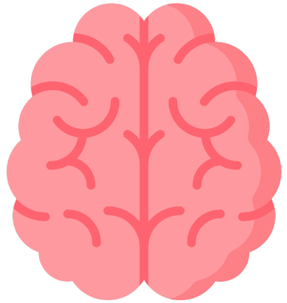
Outline

Outline

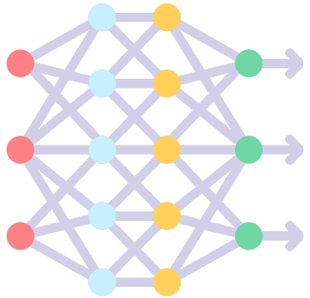


Cognitive science

Outline

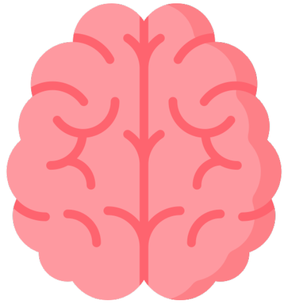


Cognitive science

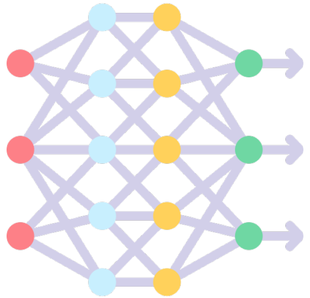


Machine learning

Outline



Cognitive science

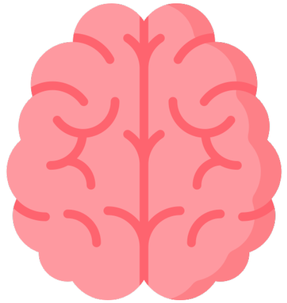


Machine learning



Large language models

Outline



Cognitive science

Machine learning



Large language models

Could I have done anything better?



Could I have done anything better?

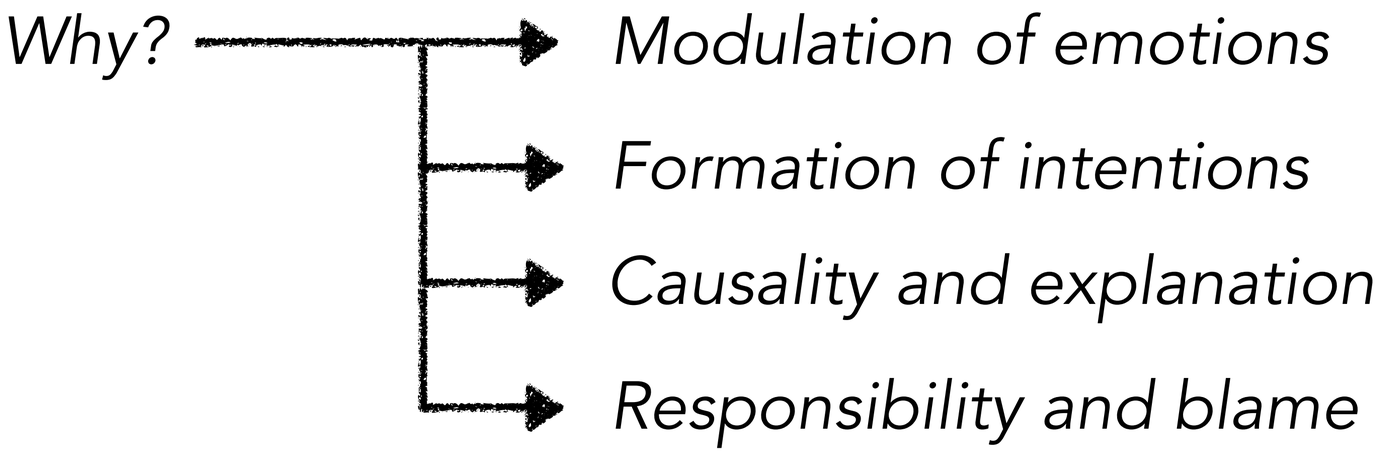




We think of counterfactuals all the time

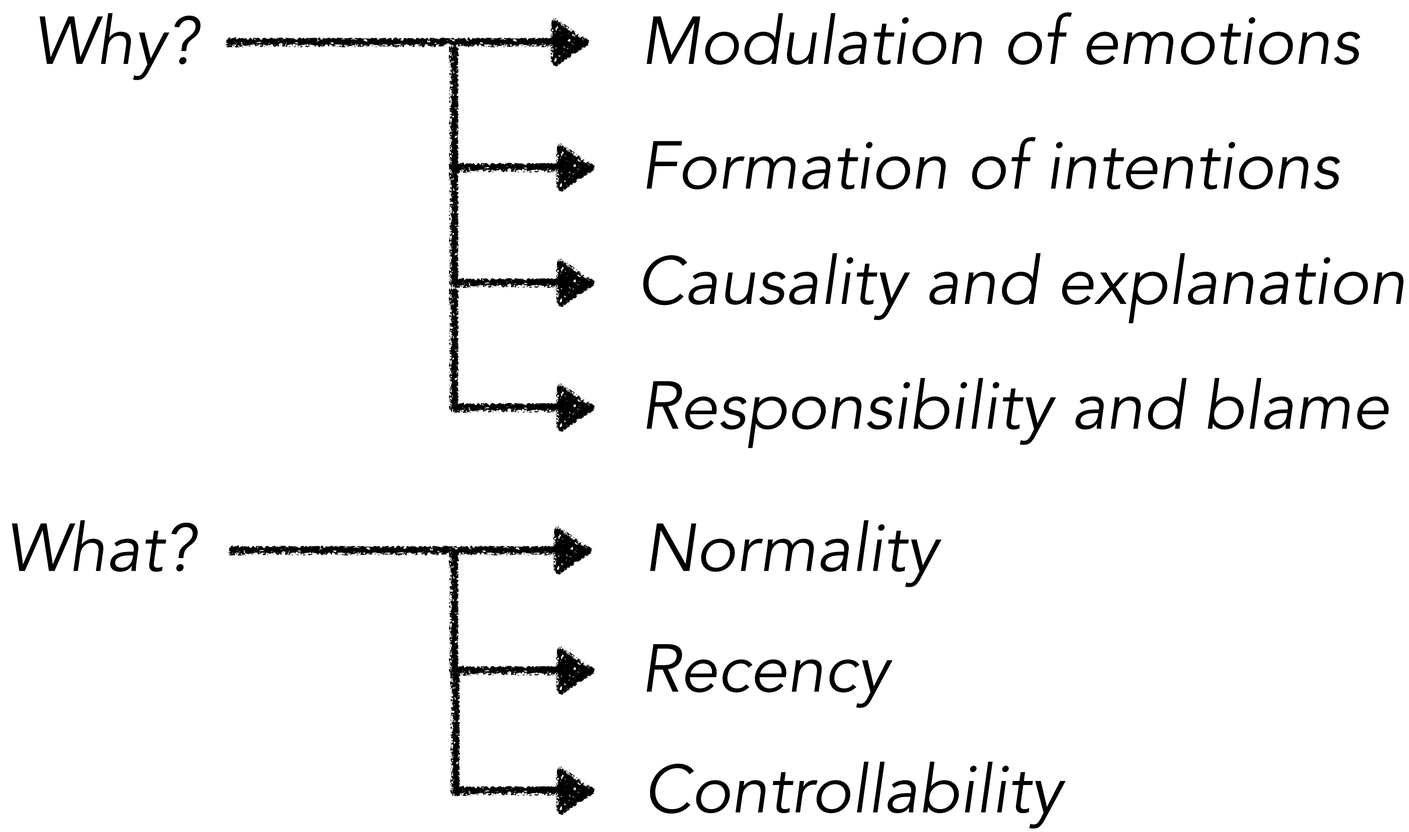
Roese. "*Counterfactual thinking.*" Psychological bulletin, 1997.
Byrne. "*Counterfactual thought.*" Annual review of psychology, 2016.

We think of counterfactuals all the time



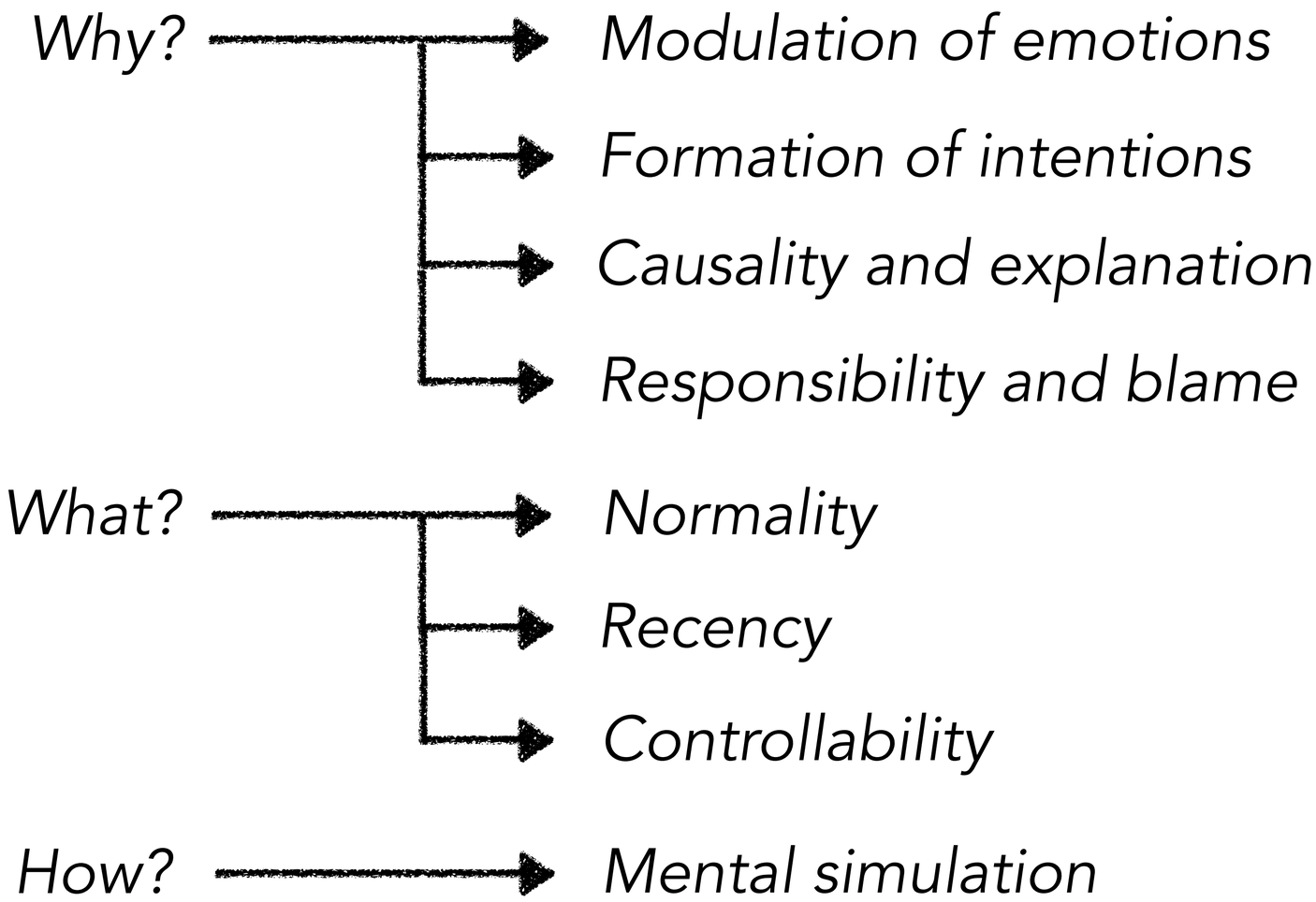
Roese. "*Counterfactual thinking.*" Psychological bulletin, 1997.
Byrne. "*Counterfactual thought.*" Annual review of psychology, 2016.

We think of counterfactuals all the time



Roese. "*Counterfactual thinking.*" Psychological bulletin, 1997.
Byrne. "*Counterfactual thought.*" Annual review of psychology, 2016.

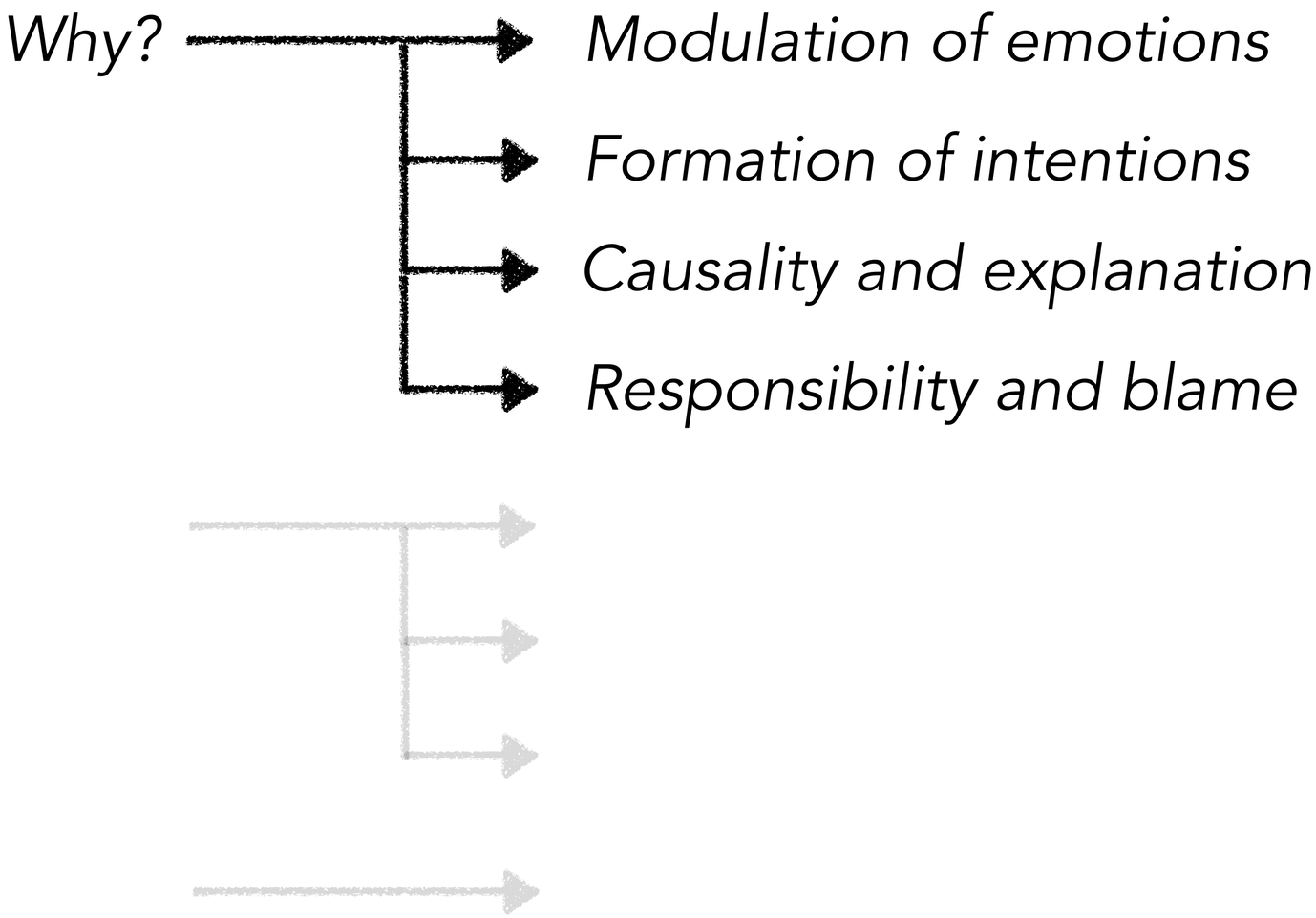
We think of counterfactuals all the time



Roese. "*Counterfactual thinking.*" Psychological bulletin, 1997.

Byrne. "*Counterfactual thought.*" Annual review of psychology, 2016.

We think of counterfactuals all the time



Roese. "*Counterfactual thinking.*" Psychological bulletin, 1997.

Byrne. "*Counterfactual thought.*" Annual review of psychology, 2016.

Upward & downward counterfactuals

Following a (typically) bad event, we tend to think in terms of counterfactuals that could have led to a **better** or **worse** outcome.

Sanna & Turley. "*Antecedents to spontaneous counterfactual thinking: effects of expectancy violation and outcome valence.*" Personality and Social Psychology Bulletin, 1996.

Upward & downward counterfactuals

Following a (typically) bad event, we tend to think in terms of counterfactuals that could have led to a **better** or **worse** outcome.



Sanna & Turley. "*Antecedents to spontaneous counterfactual thinking: effects of expectancy violation and outcome valence.*" Personality and Social Psychology Bulletin, 1996.

Upward & downward counterfactuals

Following a (typically) bad event, we tend to think in terms of counterfactuals that could have led to a **better** or **worse** outcome.



I could have won



Sanna & Turley. "*Antecedents to spontaneous counterfactual thinking: effects of expectancy violation and outcome valence.*" Personality and Social Psychology Bulletin, 1996.

Upward & downward counterfactuals

Following a (typically) bad event, we tend to think in terms of counterfactuals that could have led to a **better** or **worse** outcome.

I could have won



I could have lost faster

Sanna & Turley. "*Antecedents to spontaneous counterfactual thinking: effects of expectancy violation and outcome valence.*" Personality and Social Psychology Bulletin, 1996.

Downward counterfactuals lead to positive emotions

Downward counterfactuals lead to positive emotions



Teigen & Jensen. "*Unlucky victims or lucky survivors?*" European Psychologist, 2010.

Downward counterfactuals lead to positive emotions

Tourists who survived the 2004 tsunami were found to think 10 times more frequently about **downward** counterfactuals rather than **upward**.

I was unlucky. I could have come a week earlier.



I was lucky. I could have been severely injured.

Teigen & Jensen. "*Unlucky victims or lucky survivors?*" European Psychologist, 2010.

Upward counterfactuals lead to negative emotions

Upward counterfactuals lead to negative emotions



Medvec et al. "*When less is more: counterfactual thinking and satisfaction among Olympic medalists.*" Journal of personality and social psychology, 1995.

Upward counterfactuals lead to negative emotions

Silver medalists showed decreased happiness levels when finding out they had been second compared to bronze medalists when finding out they had been third.

I could have been first



I could have lost the medal

Medvec et al. "*When less is more: counterfactual thinking and satisfaction among Olympic medalists.*" Journal of personality and social psychology, 1995.

Upward counterfactuals lead to negative emotions

Silver medalists showed decreased happiness levels when finding out they had been second compared to bronze medalists when finding out they had been third.



I could have been first

✅ **aid self-improvement and learning from mistakes**

I could have lost the medal

Medvec et al. "*When less is more: counterfactual thinking and satisfaction among Olympic medalists.*" Journal of personality and social psychology, 1995.

McMullen & Markman. "*Downward counterfactuals and motivation: The wake-up call and the Pangloss effect.*" Personality and Social Psychology Bulletin, 2000

Formation of intentions

Reports of professional pilots after near-miss accidents were found to contain statements about upward counterfactuals followed by statements about future intentions and plans.

Morris & Moore. "*The lessons we (don't) learn: counterfactual thinking and organizational accountability after a close call.*" Administrative Science Quarterly, 2000.

Formation of intentions

Reports of professional pilots after near-miss accidents were found to contain statements about upward counterfactuals followed by statements about future intentions and plans.



If I had understood the controller's words, I wouldn't have initiated the landing attempt

Morris & Moore. "*The lessons we (don't) learn: counterfactual thinking and organizational accountability after a close call.*" Administrative Science Quarterly, 2000.

Formation of intentions

Reports of professional pilots after near-miss accidents were found to contain statements about upward counterfactuals followed by statements about future intentions and plans.



If I had understood the controller's words, I wouldn't have initiated the landing attempt





From now on, when I am given verbal instructions by a controller, I will always repeat the instructions back to be sure

Morris & Moore. "*The lessons we (don't) learn: counterfactual thinking and organizational accountability after a close call.*" Administrative Science Quarterly, 2000.

Formation of intentions

Reports of professional pilots after near-miss accidents were found to contain statements about upward counterfactuals followed by statements about future intentions and plans.

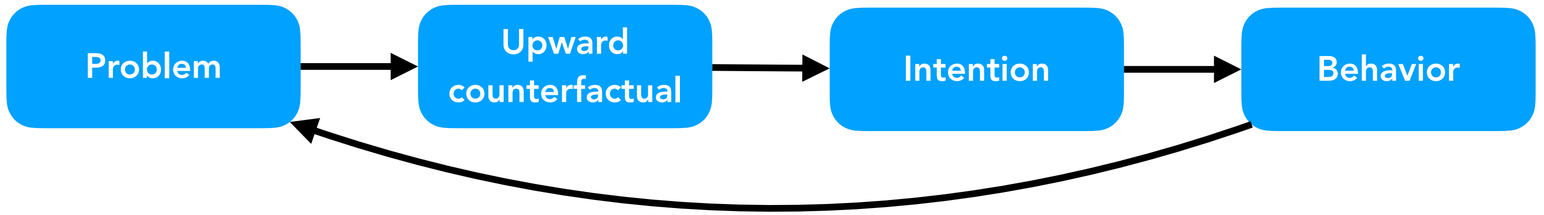


If I had understood the controller's words, I wouldn't have initiated the landing attempt





From now on, when I am given verbal instructions by a controller, I will always repeat the instructions back to be sure

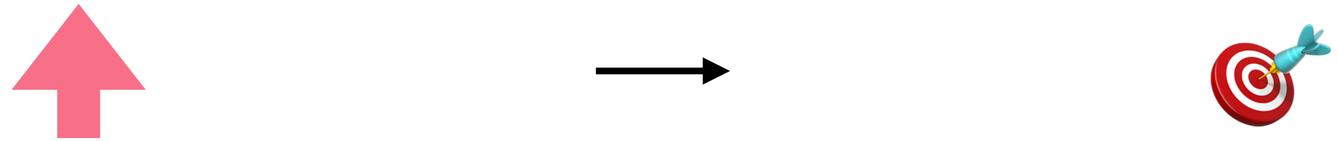


Morris & Moore. "*The lessons we (don't) learn: counterfactual thinking and organizational accountability after a close call.*" Administrative Science Quarterly, 2000.

Epstude & Roese. "*The functional theory of counterfactual thinking.*" Personality and social psychology review, 2008.

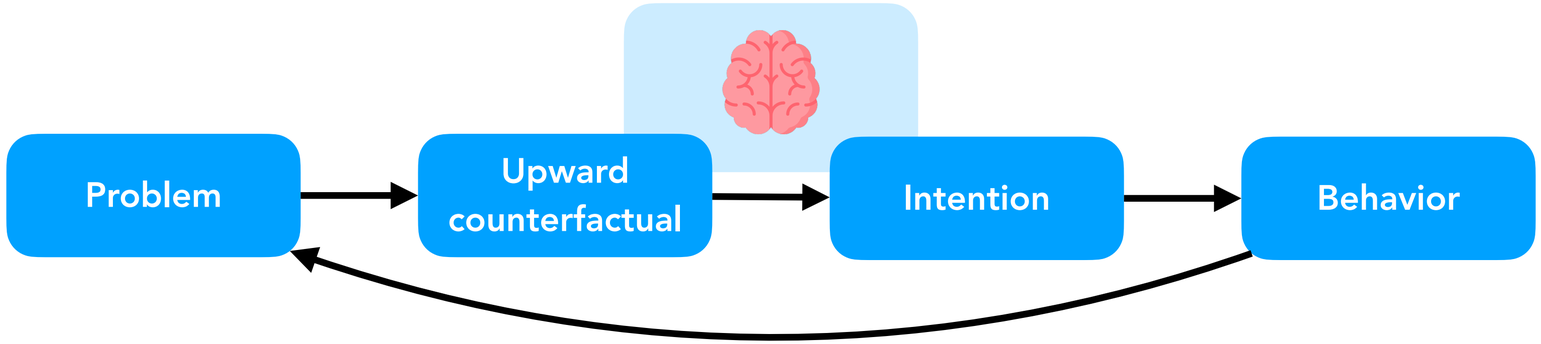
Formation of intentions

Reports of professional pilots after near-miss accidents were found to contain statements about upward counterfactuals followed by statements about future intentions and plans.



If I had understood the controller's words, I wouldn't have initiated the landing attempt

From now on, when I am given verbal instructions by a controller, I will always repeat the instructions back to be sure



Morris & Moore. "*The lessons we (don't) learn: counterfactual thinking and organizational accountability after a close call.*" Administrative Science Quarterly, 2000.

Epstude & Roese. "*The functional theory of counterfactual thinking.*" Personality and social psychology review, 2008.

Van Hoeck et al. "*Counterfactual thinking: an fMRI study on changing the past for a better future.*" Social cognitive and affective neuroscience, 2013.

Causality and explanation

Counterfactual thoughts, causal judgments and explanations of individual events have been tightly linked for many years in philosophy and psychology.

Lewis. "*Causation.*" J. Philos., 1973.

Hilton. "*Conversational processes and causal explanation.*" Psychological Bulletin, 1990.

Woodward. "*Making things happen: A theory of causal explanation.*" Oxford University Press, 2003.

Causality and explanation

Counterfactual thoughts, causal judgments and explanations of individual events have been tightly linked for many years in philosophy and psychology.

Why were you late this morning?

Lewis. "*Causation.*" J. Philos., 1973.

Hilton. "*Conversational processes and causal explanation.*" Psychological Bulletin, 1990.

Woodward. "*Making things happen: A theory of causal explanation.*" Oxford University Press, 2003.

Causality and explanation

Counterfactual thoughts, causal judgments and explanations of individual events have been tightly linked for many years in philosophy and psychology.

Why were you late this morning? → **Because** I missed the bus

Lewis. "*Causation.*" J. Philos., 1973.

Hilton. "*Conversational processes and causal explanation.*" Psychological Bulletin, 1990.

Woodward. "*Making things happen: A theory of causal explanation.*" Oxford University Press, 2003.

Causality and explanation

Counterfactual thoughts, causal judgments and explanations of individual events have been tightly linked for many years in philosophy and psychology.

Why were you late this morning? → **Because** I missed the bus → **Had I not missed the bus**, I would have been on time

Lewis. "*Causation.*" J. Philos., 1973.

Hilton. "*Conversational processes and causal explanation.*" Psychological Bulletin, 1990.

Woodward. "*Making things happen: A theory of causal explanation.*" Oxford University Press, 2003.

Causality and explanation

Counterfactual thoughts, causal judgments and explanations of individual events have been tightly linked for many years in philosophy and psychology.

Why were you late this morning? → **Because** I missed the bus → **Had I not missed the bus**, I would have been on time

Explanation = Identification of causes + Communication

Lewis. "*Causation.*" J. Philos., 1973.

Hilton. "*Conversational processes and causal explanation.*" Psychological Bulletin, 1990.

Woodward. "*Making things happen: A theory of causal explanation.*" Oxford University Press, 2003.

Causality and explanation

Counterfactual thoughts, causal judgments and explanations of individual events have been tightly linked for many years in philosophy and psychology.

Why were you late this morning? → **Because** I missed the bus → **Had I not missed the bus**, I would have been on time

Explanation = **Identification of causes** + Communication

(counterfactuals are used for this)

Lewis. "*Causation.*" J. Philos., 1973.

Hilton. "*Conversational processes and causal explanation.*" Psychological Bulletin, 1990.

Woodward. "*Making things happen: A theory of causal explanation.*" Oxford University Press, 2003.

Responsibility and blame

It is common practice for lawyers to use “but for” arguments to determine a defendant’s responsibility by establishing a causal relationship between their actions and the outcome.



Hart and Honoré. “Causation in the Law”. Oxford University Press, 1985.

Lagnado et al. "*Causal responsibility and counterfactuals.*" Cognitive science, 2013.

Responsibility and blame

It is common practice for lawyers to use “but for” arguments to determine a defendant’s responsibility by establishing a causal relationship between their actions and the outcome.

Are causality, responsibility, and blame all the same thing?



Hart and Honoré. “Causation in the Law”. Oxford University Press, 1985.

Lagnado et al. "*Causal responsibility and counterfactuals.*" Cognitive science, 2013.

Responsibility and blame

It is common practice for lawyers to use “but for” arguments to determine a defendant’s responsibility by establishing a causal relationship between their actions and the outcome.

Are causality, responsibility, and blame all the same thing?



When a drug prescription harms a patient, people hold the doctor more responsible when there is a better alternative.

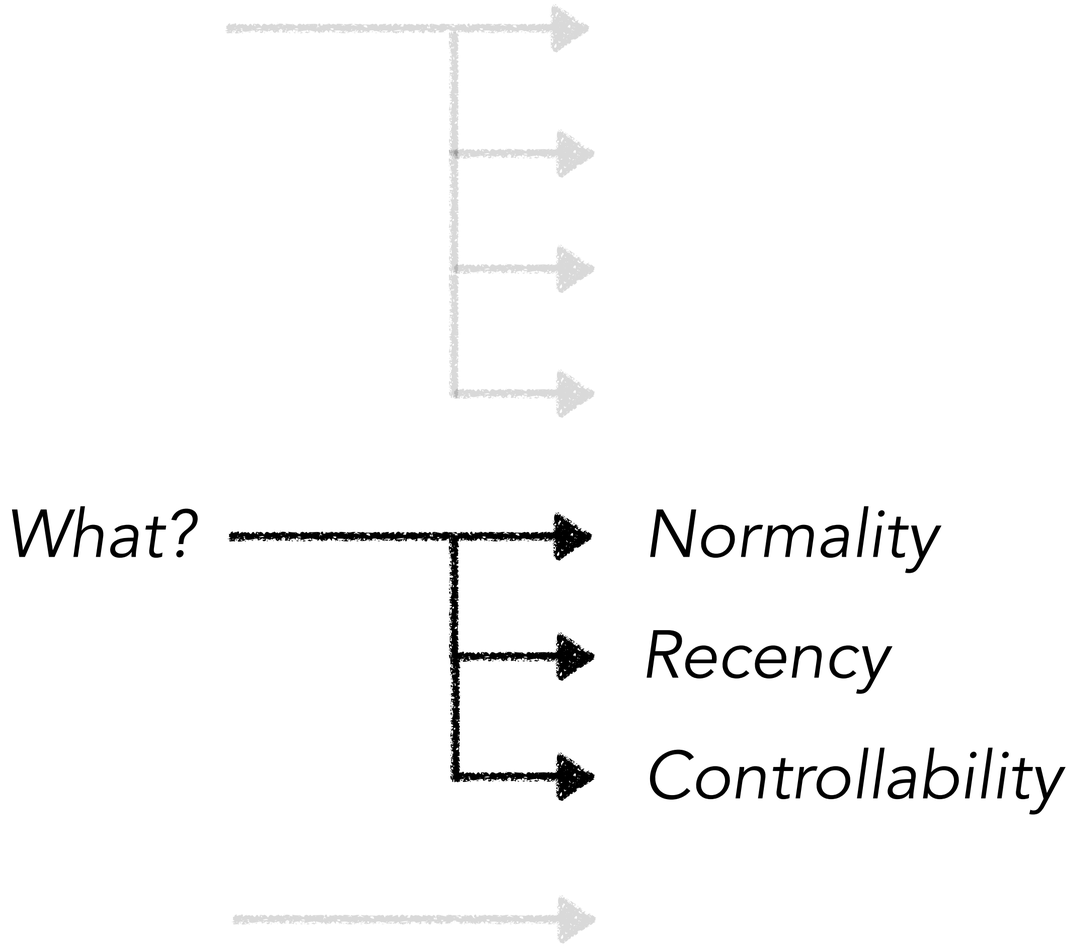
Hart and Honoré. “Causation in the Law”. Oxford University Press, 1985.

Lagnado et al. "*Causal responsibility and counterfactuals.*" Cognitive science, 2013.

Malle et al. “*A theory of blame.*” Psychological Inquiry, 2014.

Alicke et al. "*Culpable control and counterfactual reasoning in the psychology of blame.*" Personality and Social Psychology Bulletin, 2008.

We think of counterfactuals all the time



Roese. "*Counterfactual thinking.*" Psychological bulletin, 1997.
Byrne. "*Counterfactual thought.*" Annual review of psychology, 2016.

Factors that affect the choice of counterfactual contrasts

Factors that affect the choice of counterfactual contrasts

Normality

Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Factors that affect the choice of counterfactual contrasts

Normality

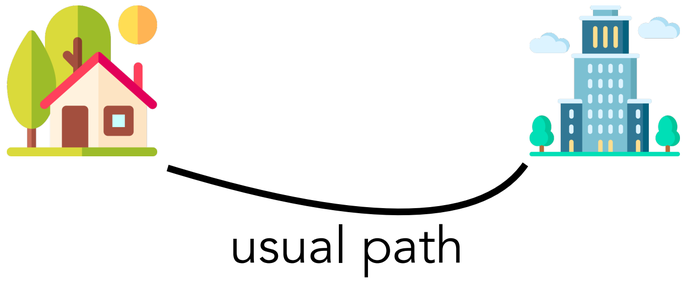




Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

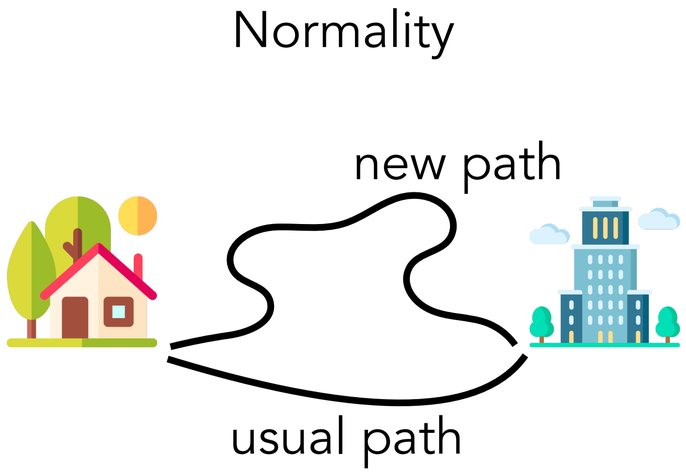
Factors that affect the choice of counterfactual contrasts

Normality



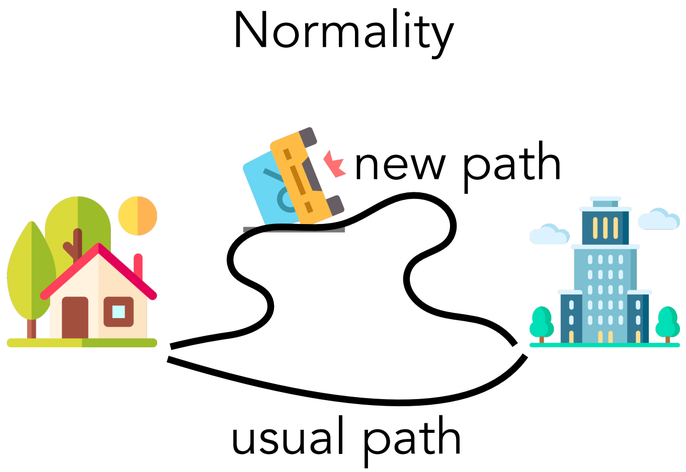
Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Factors that affect the choice of counterfactual contrasts



Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Factors that affect the choice of counterfactual contrasts



Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Factors that affect the choice of counterfactual contrasts

Normality



if-likelihood × then-likelihood

Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Petrocelli et al. "*Counterfactual potency.*" Journal of personality and social psychology, 2011.

Factors that affect the choice of counterfactual contrasts

Normality



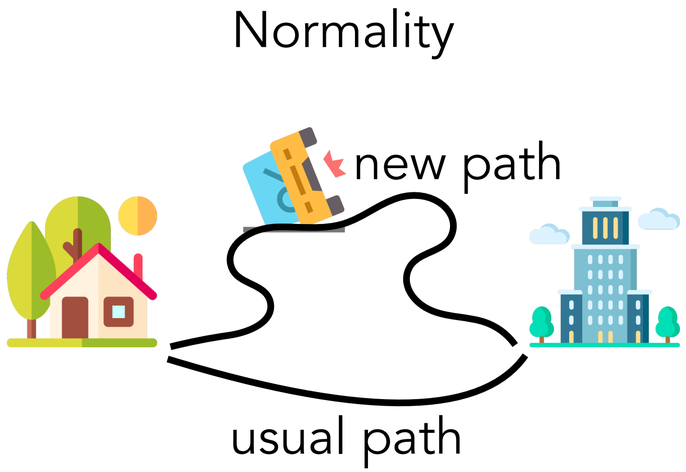
if-likelihood × **then-likelihood**

hypothetical

Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Petrocelli et al. "*Counterfactual potency.*" Journal of personality and social psychology, 2011.

Factors that affect the choice of counterfactual contrasts



if-likelihood × **then-likelihood**

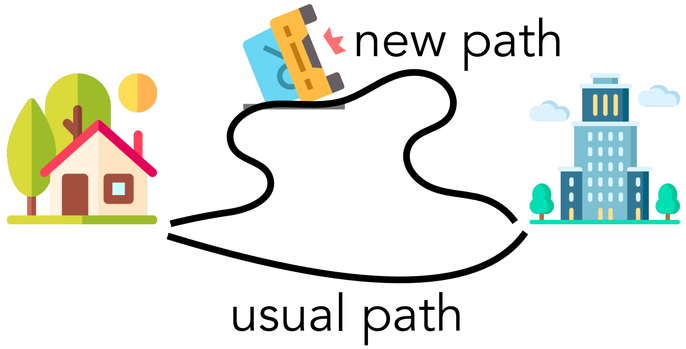
hypothetical **counterfactual**

Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Petrocelli et al. "*Counterfactual potency.*" Journal of personality and social psychology, 2011.

Factors that affect the choice of counterfactual contrasts

Normality



if-likelihood × **then-likelihood**

hypothetical **counterfactual**

Recency



Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

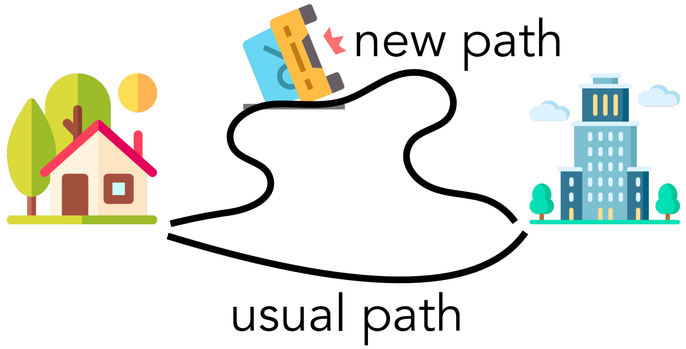
Petrocelli et al. "*Counterfactual potency.*" Journal of personality and social psychology, 2011.

Spellman. "*Crediting causality.*" Journal of Experimental Psychology, 1997.

Gerstenberg and Lagnado. "*When contributions make a difference: Explaining order effects in responsibility attribution.*" Psychonomic bulletin & review, 2012

Factors that affect the choice of counterfactual contrasts

Normality



if-likelihood × then-likelihood

hypothetical counterfactual

Recency



Controllability

Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Petrocelli et al. "*Counterfactual potency.*" Journal of personality and social psychology, 2011.

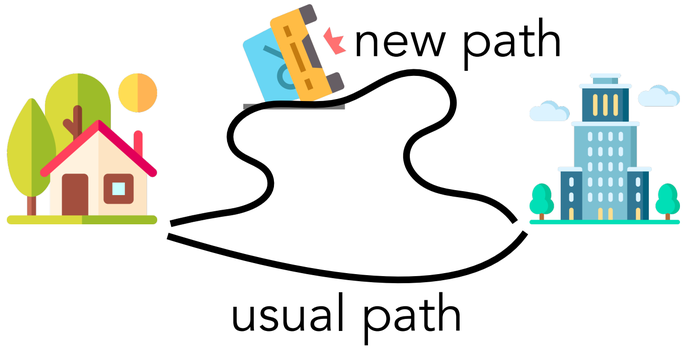
Spellman. "*Crediting causality.*" Journal of Experimental Psychology, 1997.

Gerstenberg and Lagnado. "*When contributions make a difference: Explaining order effects in responsibility attribution.*" Psychonomic bulletin & review, 2012

Girotto et al. "*Event controllability in counterfactual thinking.*" Acta Psychologica, 1991.

Factors that affect the choice of counterfactual contrasts

Normality



if-likelihood × **then-likelihood**

hypothetical **counterfactual**

Recency



Controllability



Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

Petrocelli et al. "*Counterfactual potency.*" Journal of personality and social psychology, 2011.

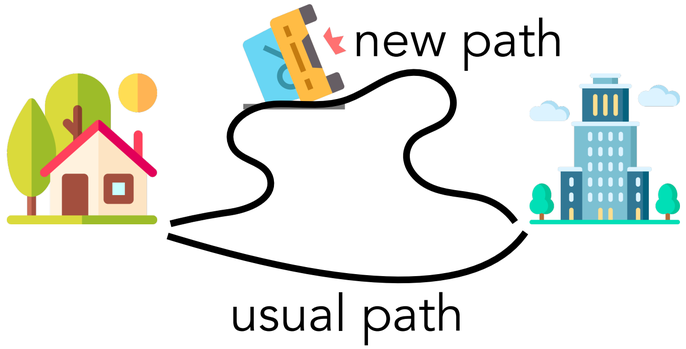
Spellman. "*Crediting causality.*" Journal of Experimental Psychology, 1997.

Gerstenberg and Lagnado. "*When contributions make a difference: Explaining order effects in responsibility attribution.*" Psychonomic bulletin & review, 2012

Girotto et al. "*Event controllability in counterfactual thinking.*" Acta Psychologica, 1991.

Factors that affect the choice of counterfactual contrasts

Normality



if-likelihood × **then-likelihood**

hypothetical **counterfactual**

Recency



Controllability



Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" Psychological review, 1986.

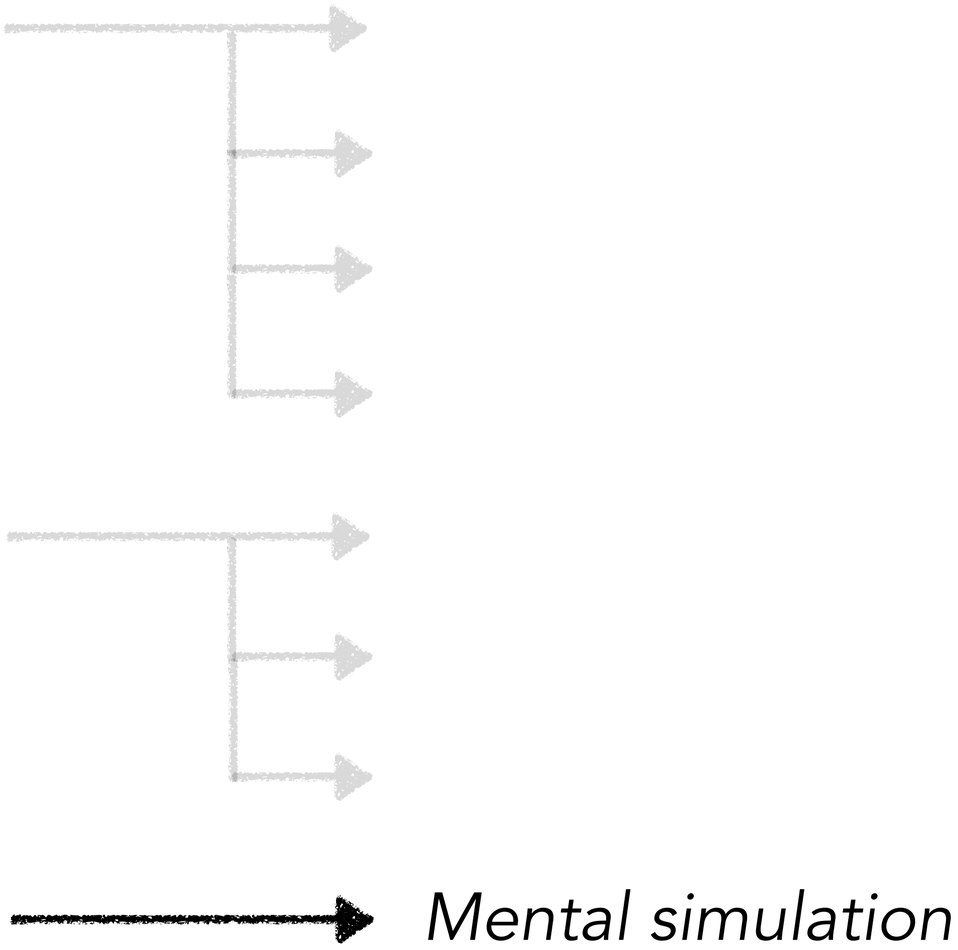
Petrocelli et al. "*Counterfactual potency.*" Journal of personality and social psychology, 2011.

Spellman. "*Crediting causality.*" Journal of Experimental Psychology, 1997.

Gerstenberg and Lagnado. "*When contributions make a difference: Explaining order effects in responsibility attribution.*" Psychonomic bulletin & review, 2012

Girotto et al. "*Event controllability in counterfactual thinking.*" Acta Psychologica, 1991.

We think of counterfactuals all the time



Roese. "*Counterfactual thinking.*" Psychological bulletin, 1997.
Byrne. "*Counterfactual thought.*" Annual review of psychology, 2016.

Could I have done anything better?





No oracle available in the real world!

Mental simulation

“If the organism carries a **"small-scale model" of external reality and of its own possible actions within its head**, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, ...”

Craik (**1943**) The nature of explanation.

Mental simulation

"If the organism carries a **"small-scale model" of external reality and of its own possible actions within its head**, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, ..."

Craik (**1943**) The nature of explanation.

14. The simulation heuristic

Daniel Kahneman and Amos Tversky

Our original treatment of the availability heuristic (Tversky & Kahneman, 1973, **11**) discussed two classes of mental operations that "bring things to mind": the retrieval of instances and the construction of examples or scenarios. *Recall* and *construction* are quite different ways of bringing things to mind; they are used to answer different questions, and they follow different rules. Past research has dealt mainly with the retrieval of instances from memory, and the process of mental construction has been relatively neglected.

To advance the study of availability for construction, we now sketch a mental operation that we label the simulation heuristic. Our starting point is a common introspection: There appear to be many situations in which questions about events are answered by an operation that resembles the running of a simulation model. The simulation can be constrained and controlled in several ways: The starting conditions for a "run" can be left at their realistic default values or modified to assume some special

Mental simulation

"If the organism carries a **"small-scale model" of external reality and of its own possible actions within its head**, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, ..."

Craik (**1943**) The nature of explanation.

14. The simulation heuristic

Daniel Kahneman and Amos Tversky

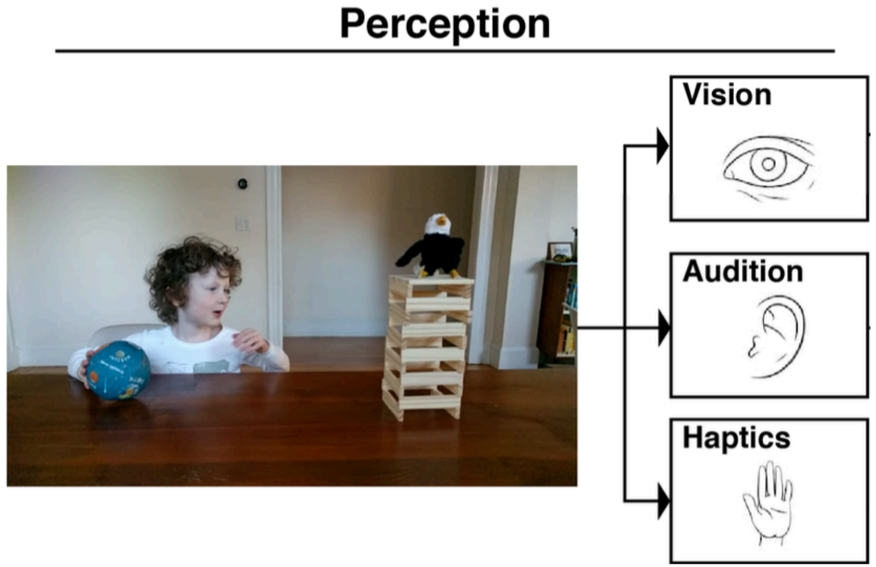
Our original treatment of the availability heuristic (Tversky & Kahneman, 1973, **11**) discussed two classes of mental operations that "bring things to mind": the retrieval of instances and the construction of examples or scenarios. *Recall* and *construction* are quite different ways of bringing things to mind; they are used to answer different questions, and they follow different rules. Past research has dealt mainly with the retrieval of instances from memory, and the process of mental construction has been relatively neglected.

To advance the study of availability for construction, we now sketch a mental operation that we label the simulation heuristic. Our starting point is a common introspection: There appear to be many situations in which questions about events are answered by an operation that resembles the running of a simulation model. The simulation can be constrained and controlled in several ways: The starting conditions for a "run" can be left at their realistic default values or modified to assume some special

Mental machinery and operations

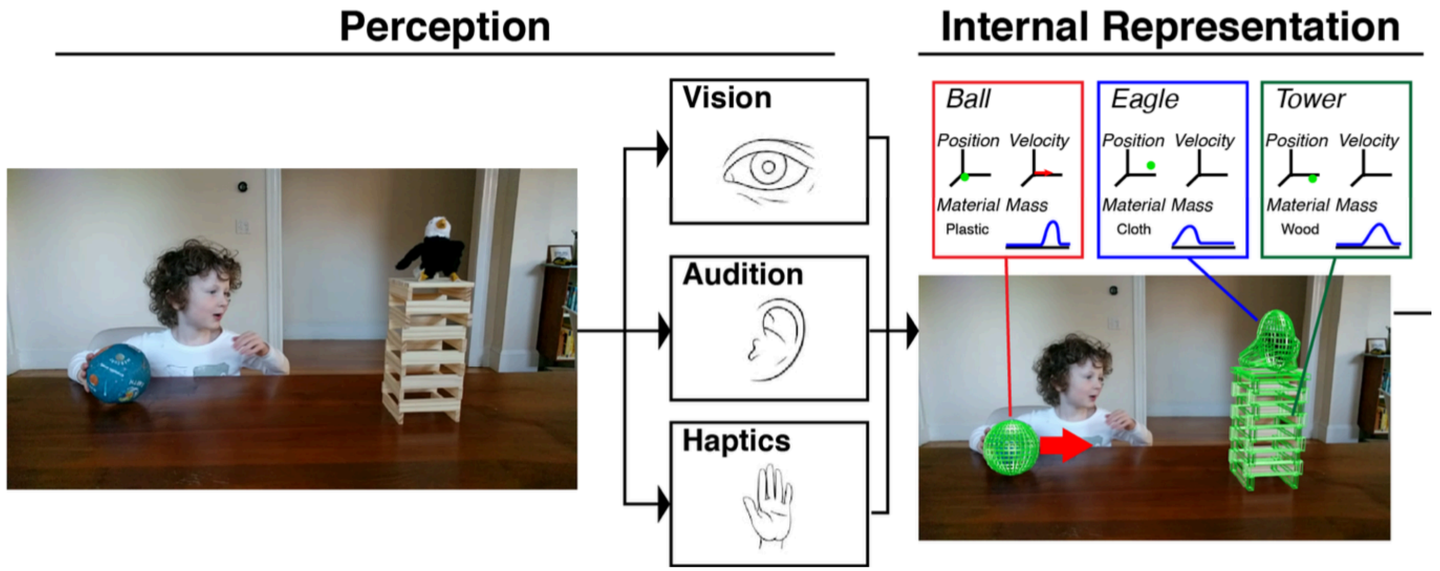
Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Mental machinery and operations



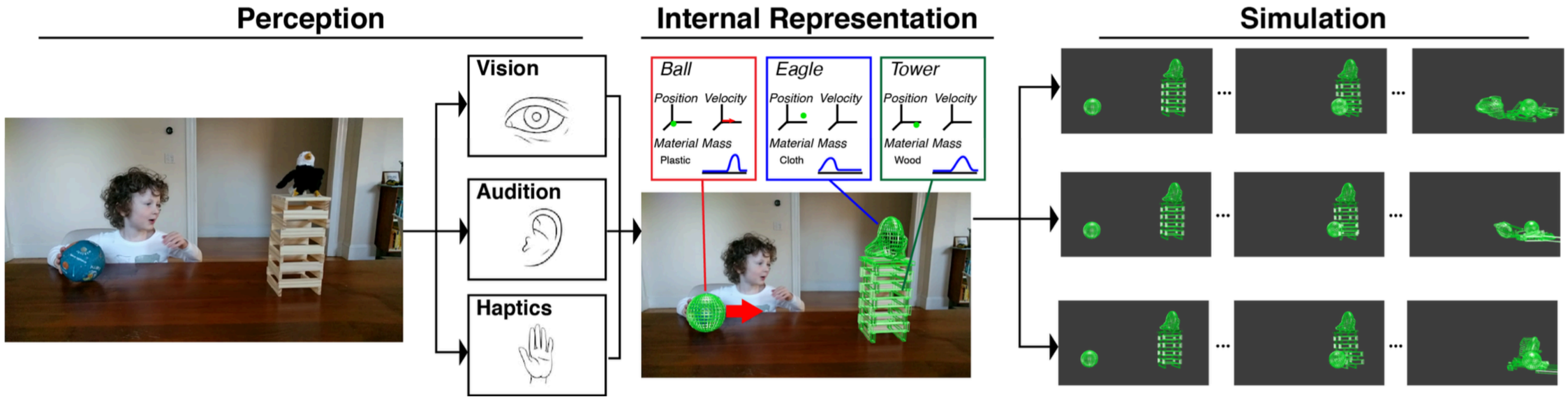
Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Mental machinery and operations



Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Mental machinery and operations



Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Goals of mental simulation



Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Goals of mental simulation



Predict what will happen

Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Goals of mental simulation

Infer what happened



Predict what will happen

Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Goals of mental simulation

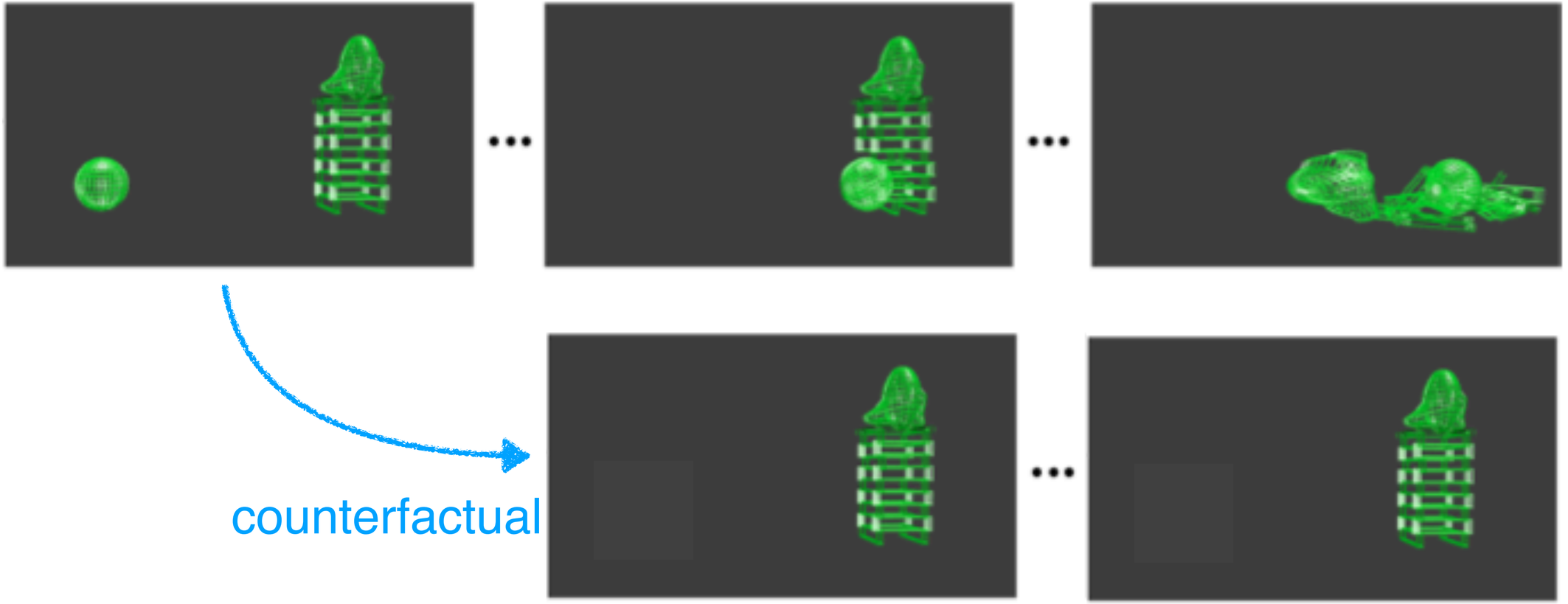
Explain why something happened



Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Goals of mental simulation

Explain why something happened



Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

Trends in
Cognitive Sciences

CellPress

Review

Counterfactual simulation in causal cognition

Tobias Gerstenberg [1,*]

How do people make causal judgments and assign responsibility? In this review article, I argue that counterfactual simulations are key. To simulate counterfactuals, we need three ingredients: a generative mental model of the world, the ability to perform interventions on that model, and the capacity to simulate the consequences of these interventions. The counterfactual simulation model (CSM) uses these ingredients to capture people's intuitive understanding of the physical and social world. In the physical domain, the CSM predicts people's causal judgments about dynamic collision events, complex situations that involve multiple causes, omissions as causes, and causes that sustain physical stability. In the social domain, the CSM predicts responsibility judgments in helping and hindering scenarios.

Highlights

People judge causation and attribute responsibility by simulating counterfactual alternatives.

The counterfactual simulation model (CSM) captures people's causal judgments about physical events and responsibility judgments about social events.

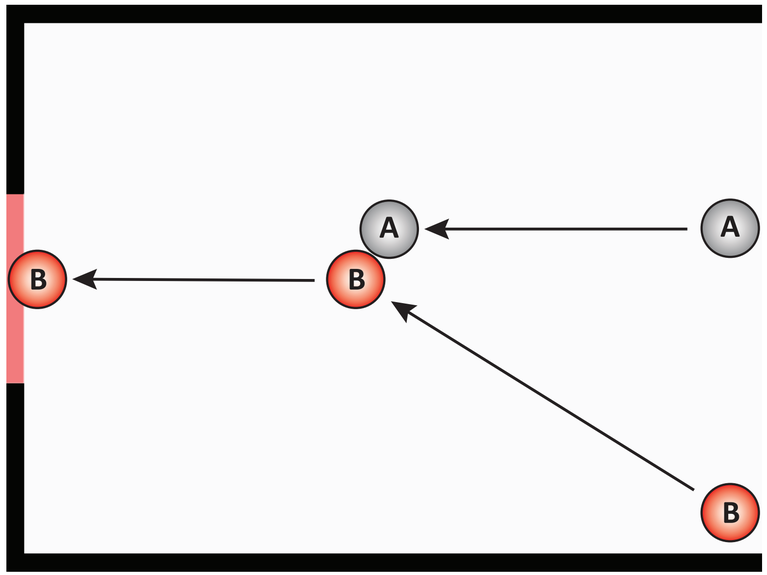
In the physical domain, the CSM pre-

Deep dive:
Counterfactual simulation for causal judgments

Gerstenberg et al. "*A counterfactual simulation model of causal judgments for physical events.*" Psychological review, 2021.

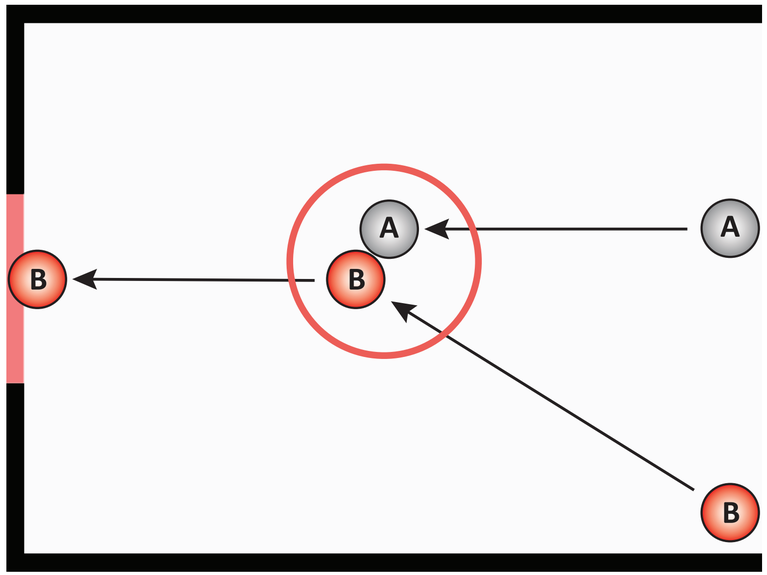
Counterfactual Simulation Model

What happened?



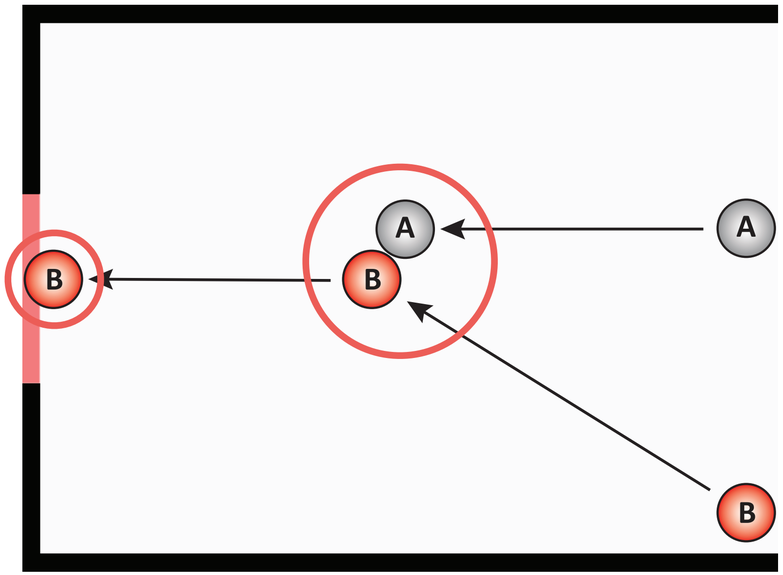
Counterfactual Simulation Model

What happened?



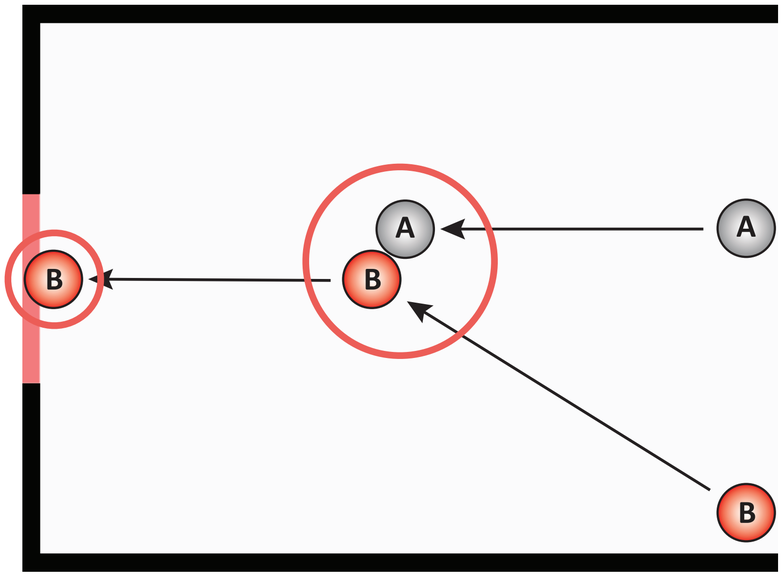
Counterfactual Simulation Model

What happened?

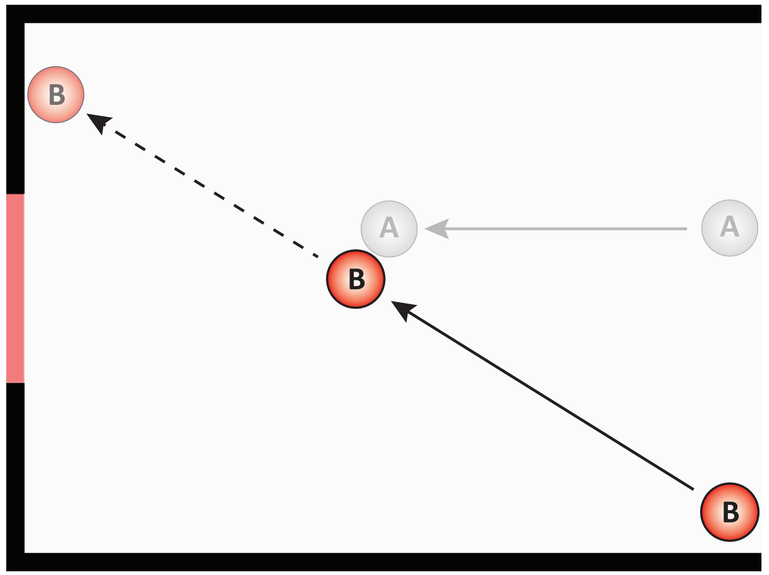


Counterfactual Simulation Model

What happened?

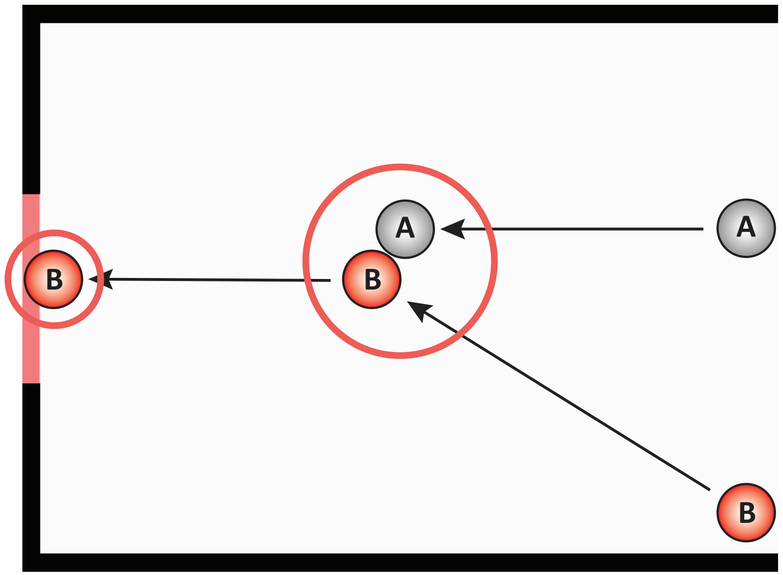


What would have happened?

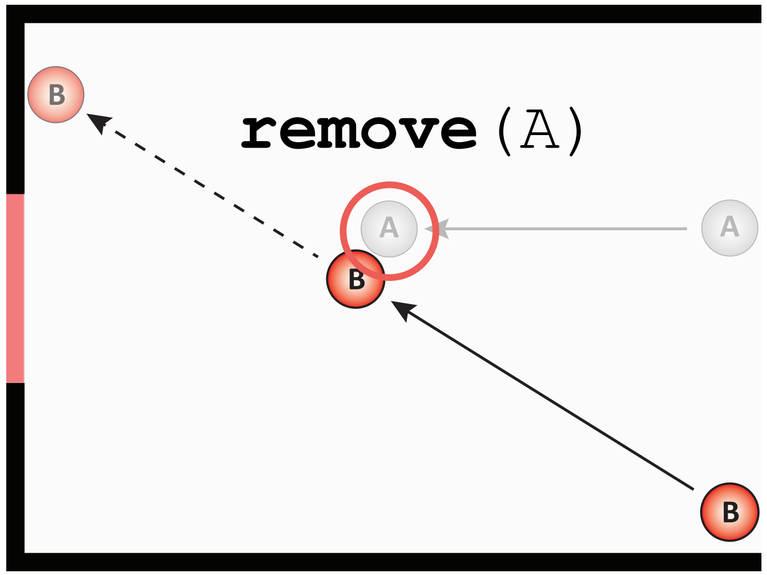


Counterfactual Simulation Model

What happened?

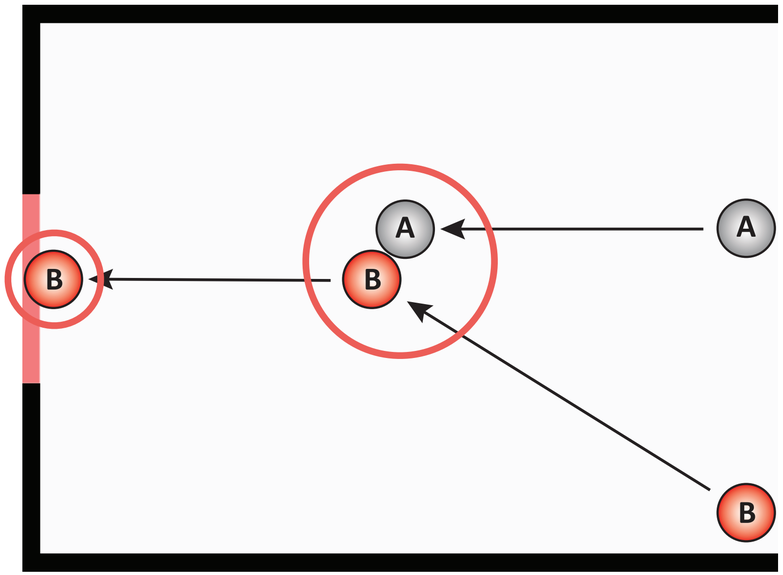


What would have happened?

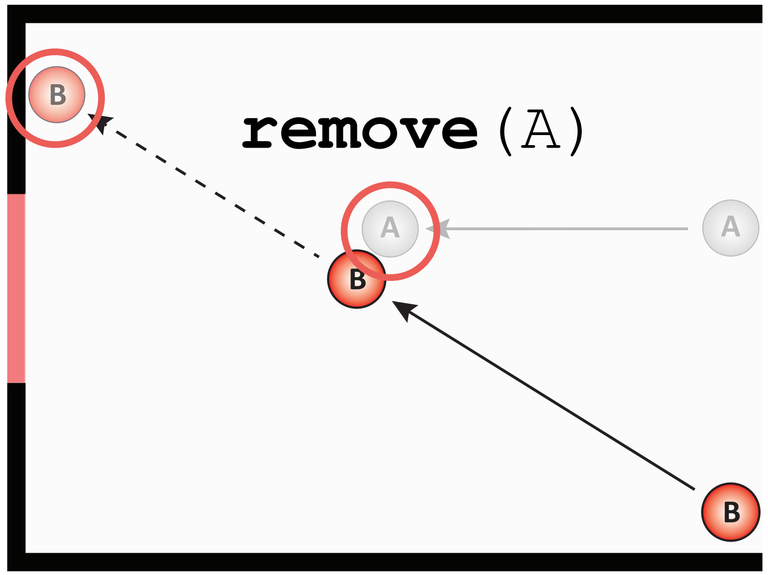


Counterfactual Simulation Model

What happened?

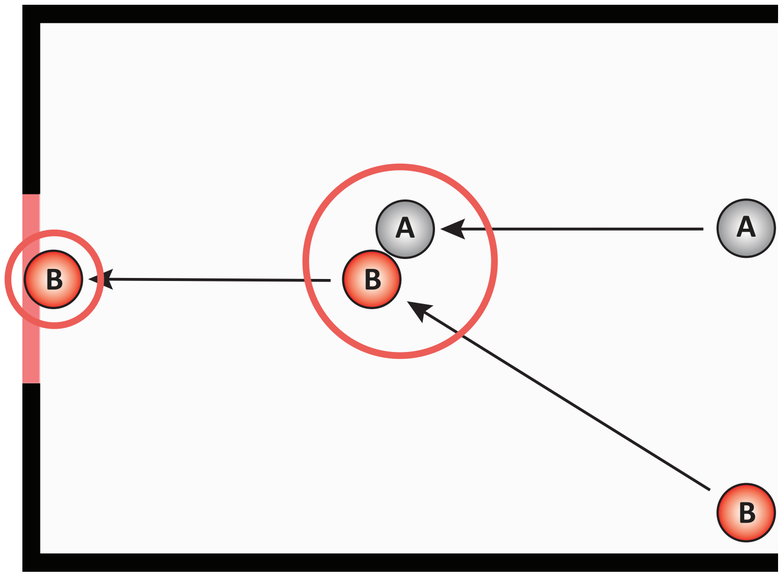


What would have happened?



Counterfactual Simulation Model

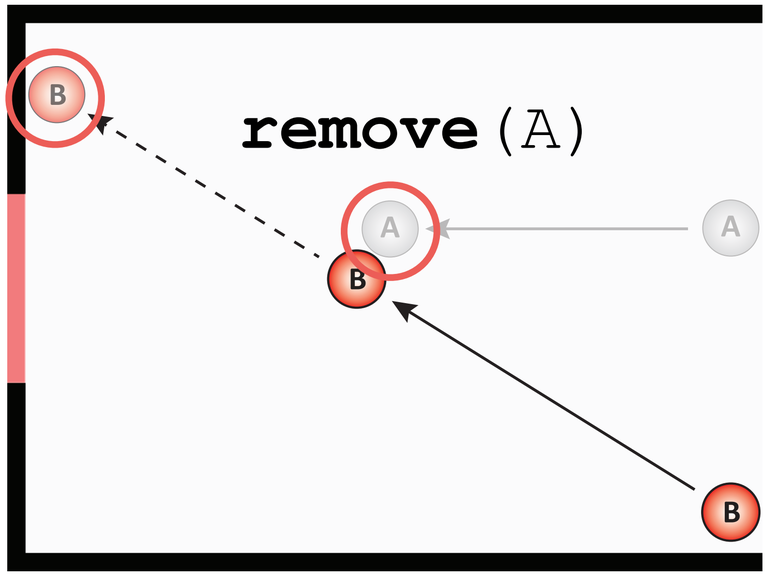
What happened?



Actual situation

B went through the gate

What would have happened?



Counterfactual situation

B would have missed the gate

Counterfactual Simulation Model

Counterfactual Simulation Model



Generative model

probabilistic program

```
//Define table with walls
function createTable(wall.x,wall.y,wall.length,wall.width){...}
//Define balls
function createBalls(x.position,y.position,x.velocity,y.velocity){...}

//Define world
function createWorld(table, ball1, ball2){
    createTable(...);
    createBalls(...);
    return(world)
}
```


Chater and Oaksford. "*Programs as causal models: Speculations on mental programs and mental representation.*" Cognitive science, 2013.

Goodman et al. "*Concepts in a probabilistic language of thought.*" In The Conceptual Mind: New Directions in the Study of Concepts, MIT Press, 2015.

Counterfactual Simulation Model



Generative model

probabilistic program

```
//Define table with walls
function createTable(wall.x,wall.y,wall.length,wall.width){...}
//Define balls
function createBalls(x.position,y.position,x.velocity,y.velocity){...}

//Define world
function createWorld(table, ball1, ball2){
    createTable(...);
    createBalls(...);
    return(world)
}
```

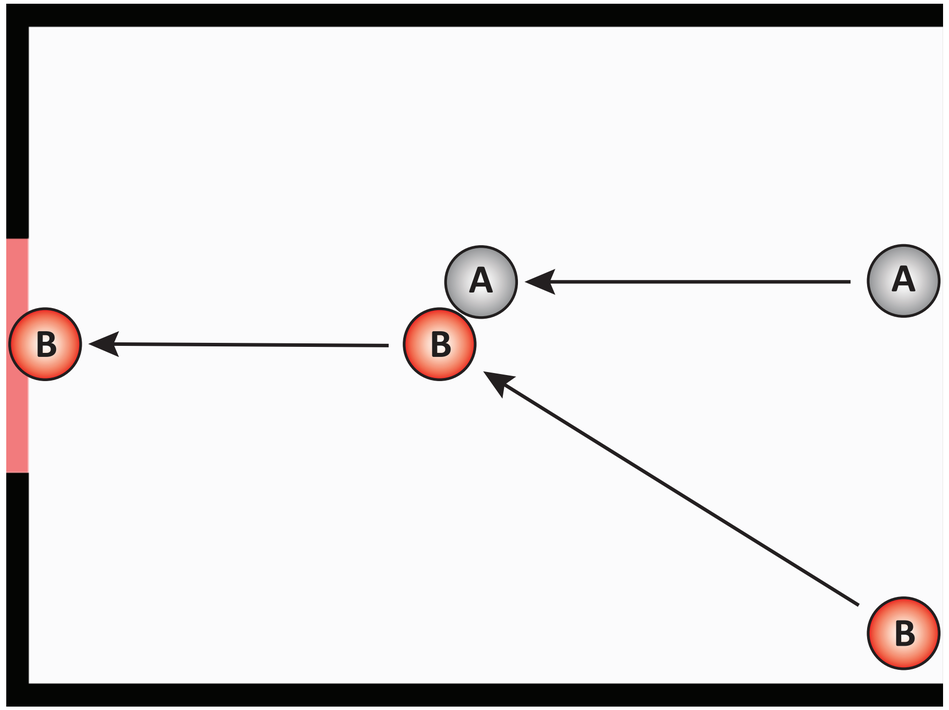
Counterfactual intervention

`remove``(object)` operator

Chater and Oaksford. "*Programs as causal models: Speculations on mental programs and mental representation.*" Cognitive science, 2013.

Goodman et al. "*Concepts in a probabilistic language of thought.*" In The Conceptual Mind: New Directions in the Study of Concepts, MIT Press, 2015.

What happened?

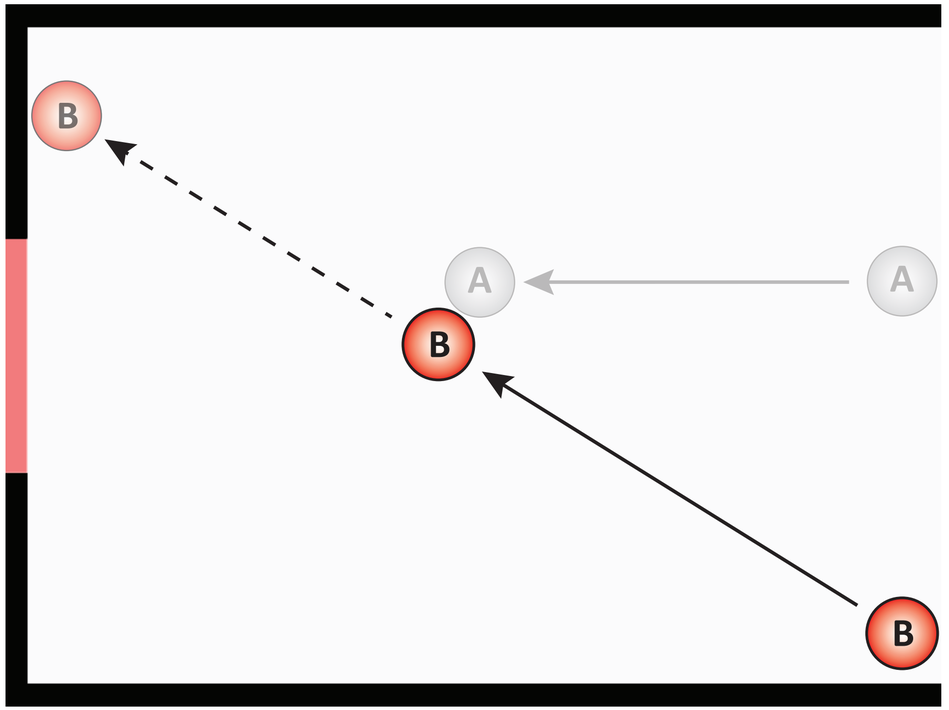


Actual situation

B went through the gate

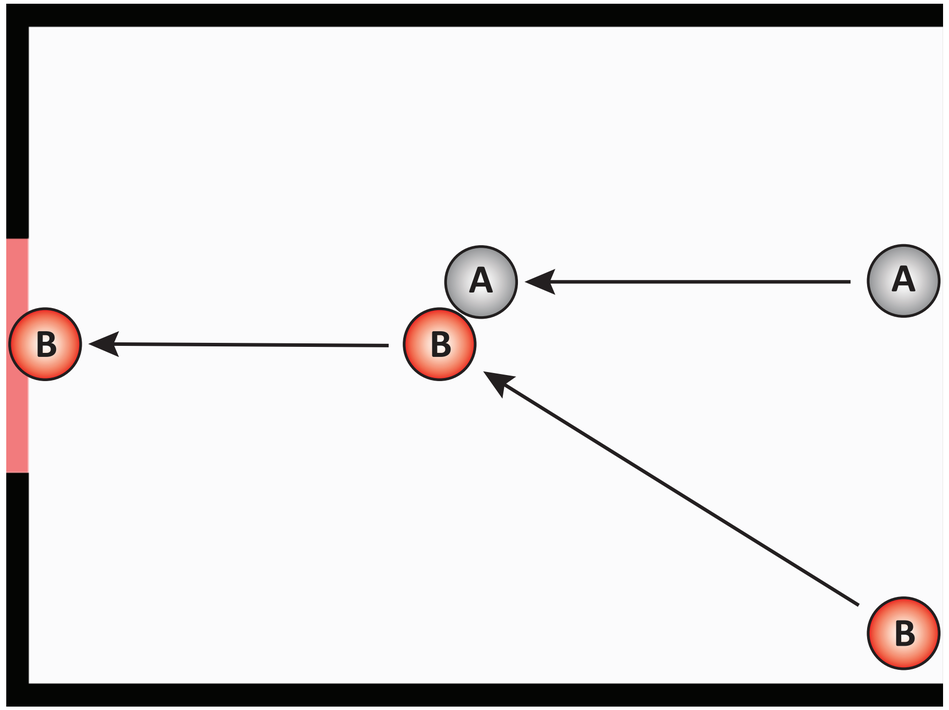
≠

What would have happened?



Counterfactual situation

What happened?

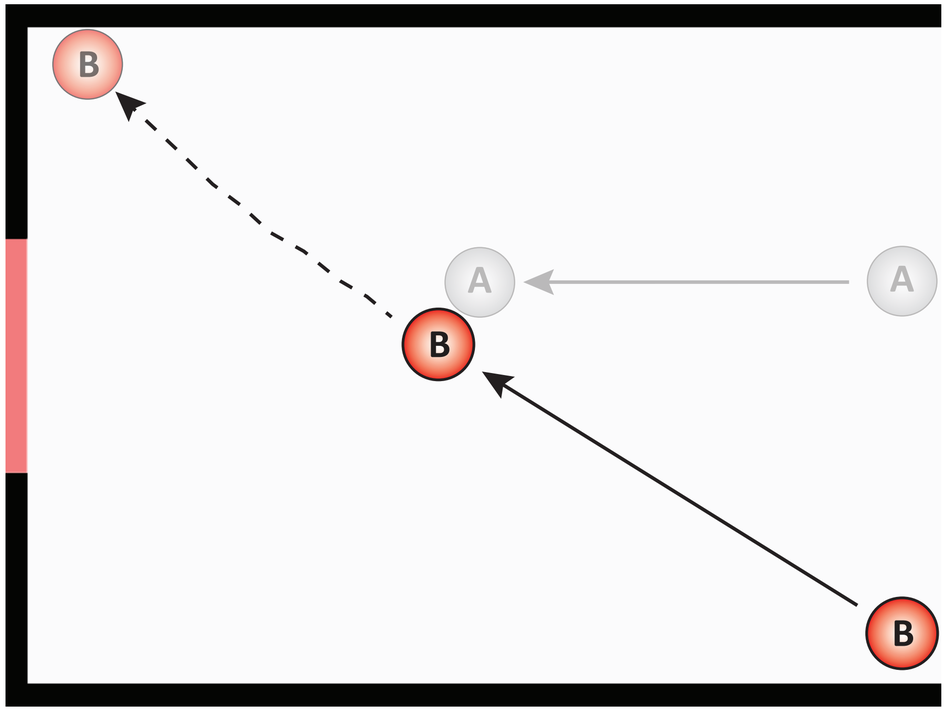


Actual situation

B went through the gate

≠

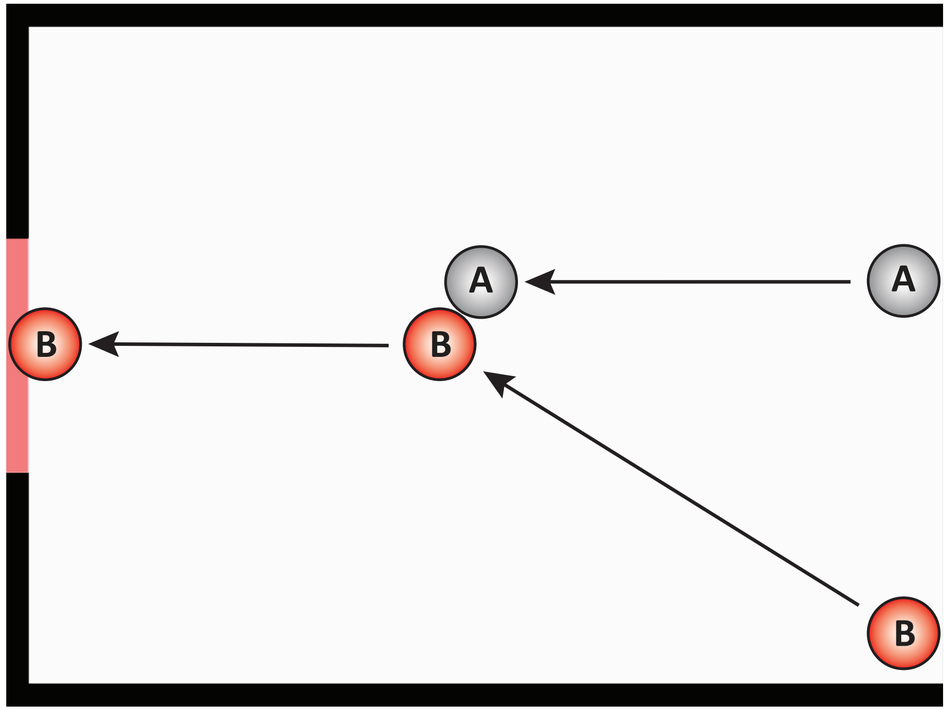
What would have happened?



Counterfactual situation

B would have missed the gate ✓

What happened?

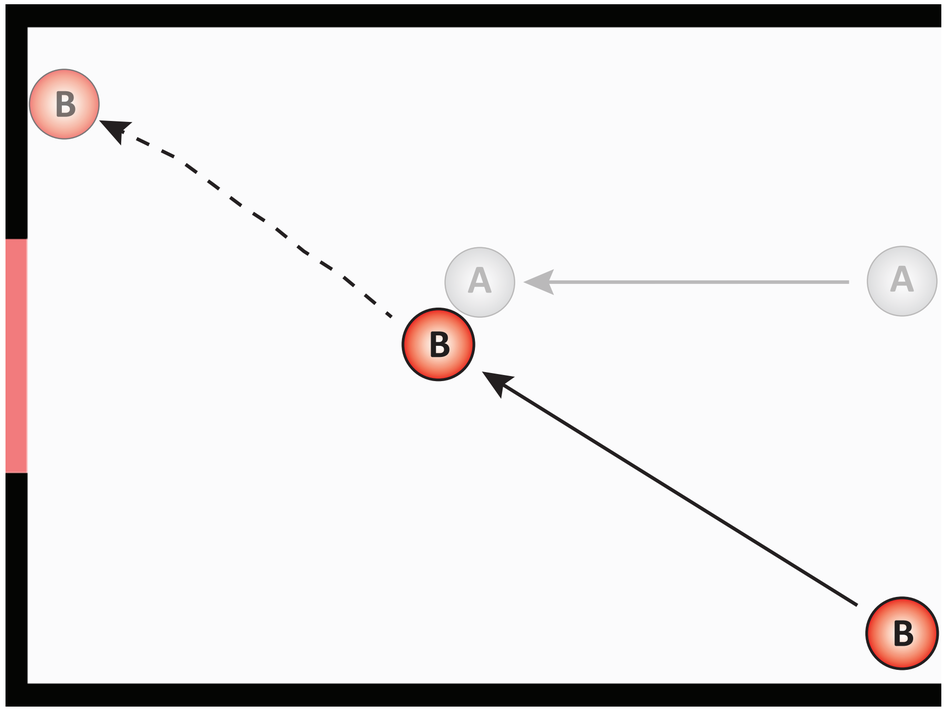


Actual situation

B went through the gate

≠

What would have happened?

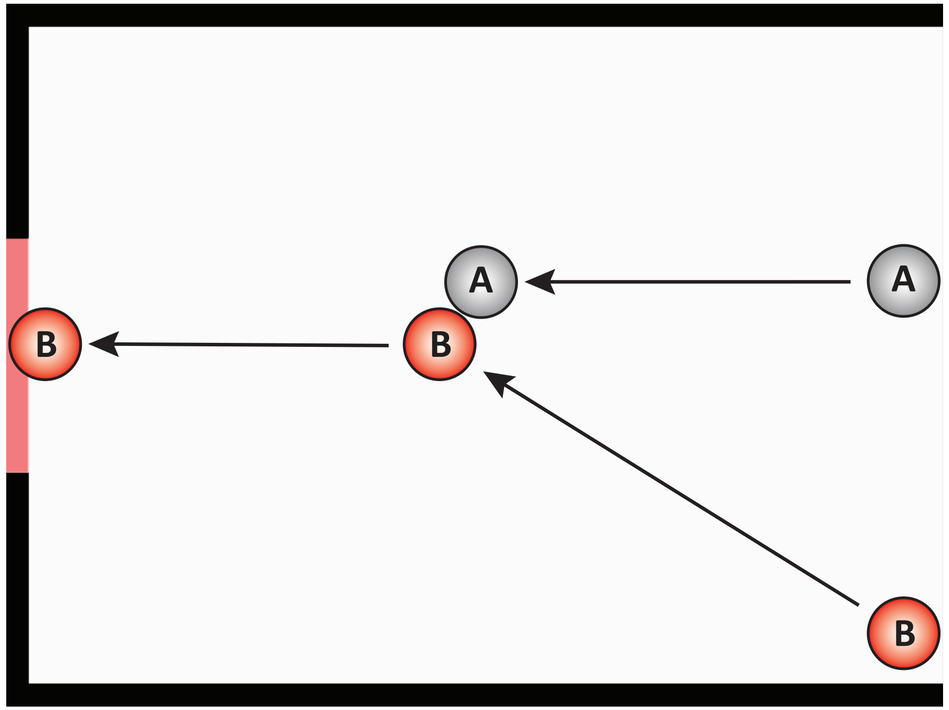


Counterfactual situation

B would have missed the gate ✓

B would have missed the gate ✓

What happened?

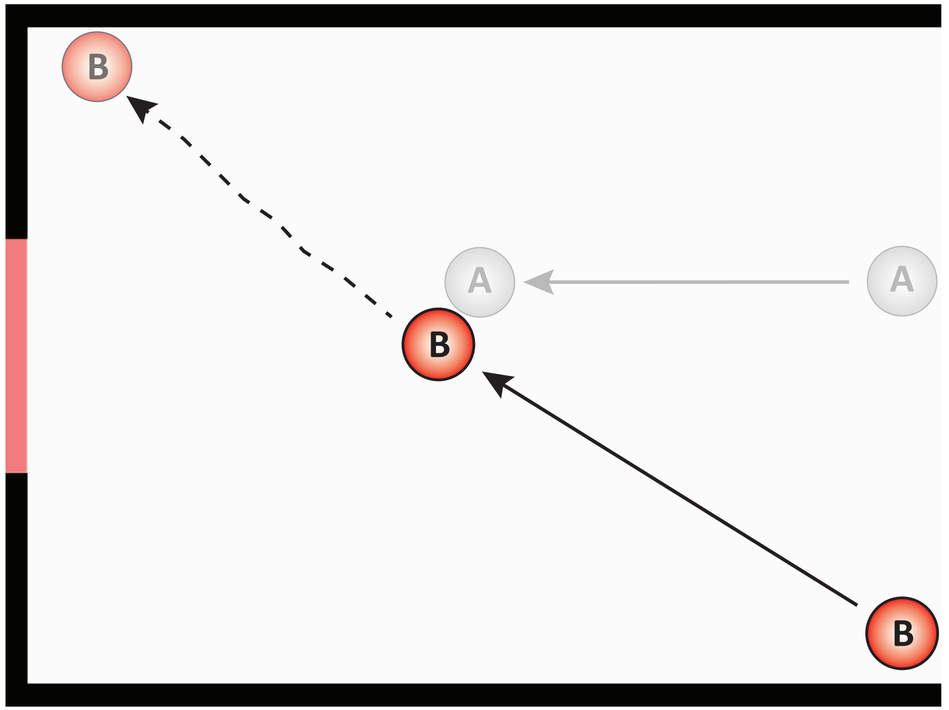


Actual situation

B went through the gate

≠

What would have happened?



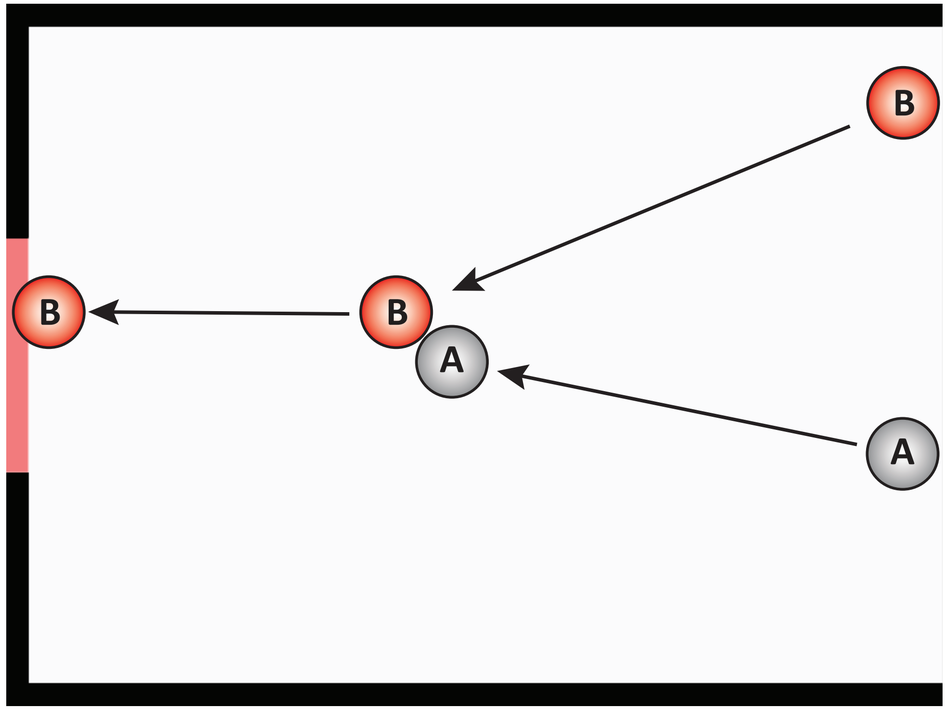
Counterfactual situation

B would have missed the gate ✓

B would have missed the gate ✓

B would have missed the gate ✓

What happened?

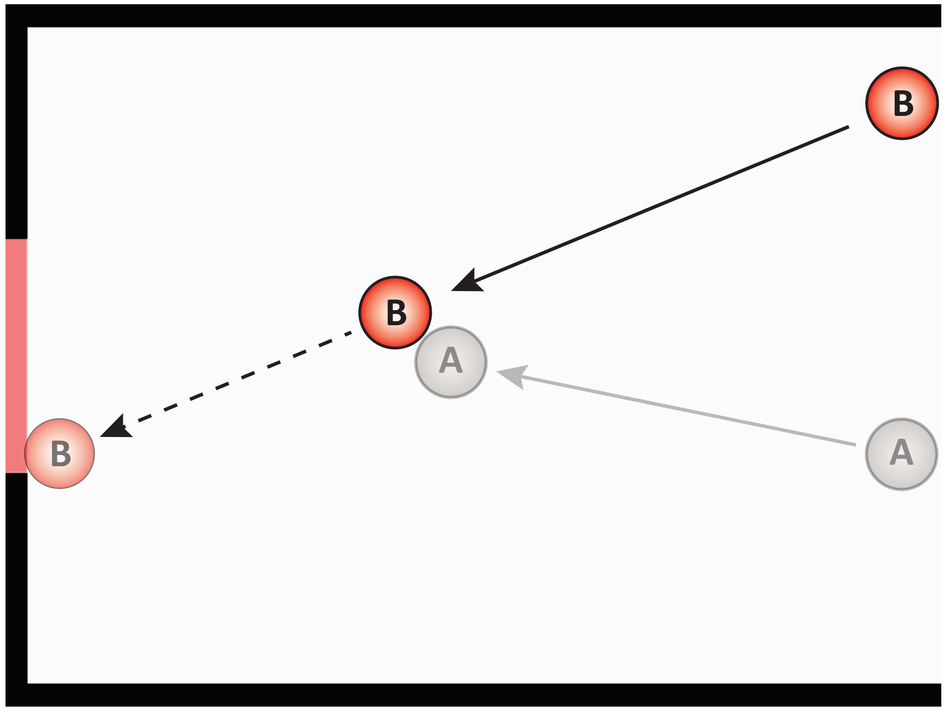


Actual situation

B went through the gate

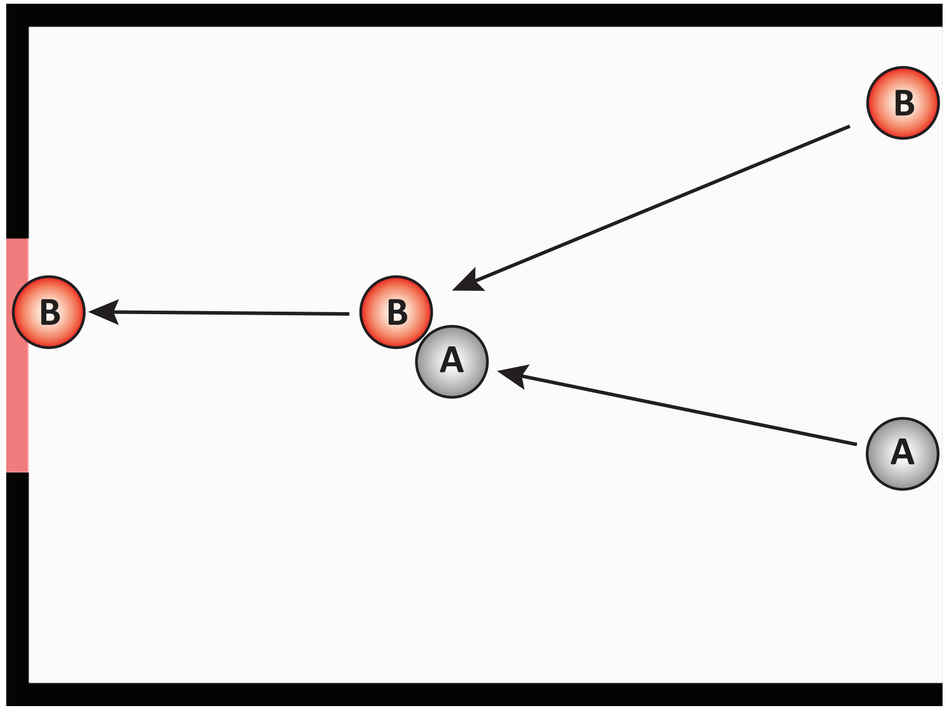
≠

What would have happened?



Counterfactual situation

What happened?

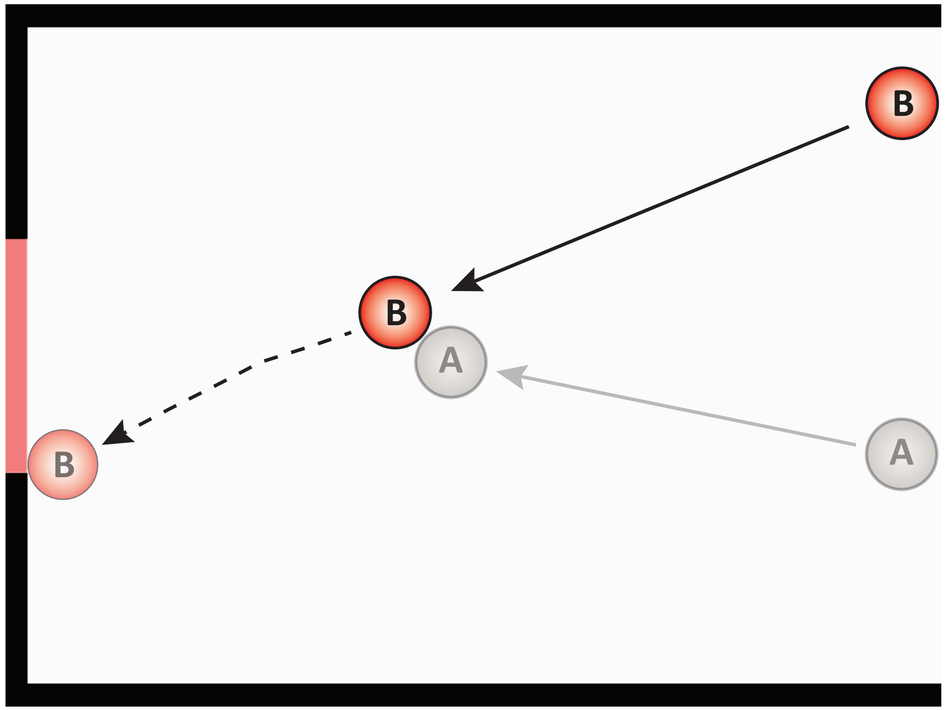


Actual situation

B went through the gate

≠

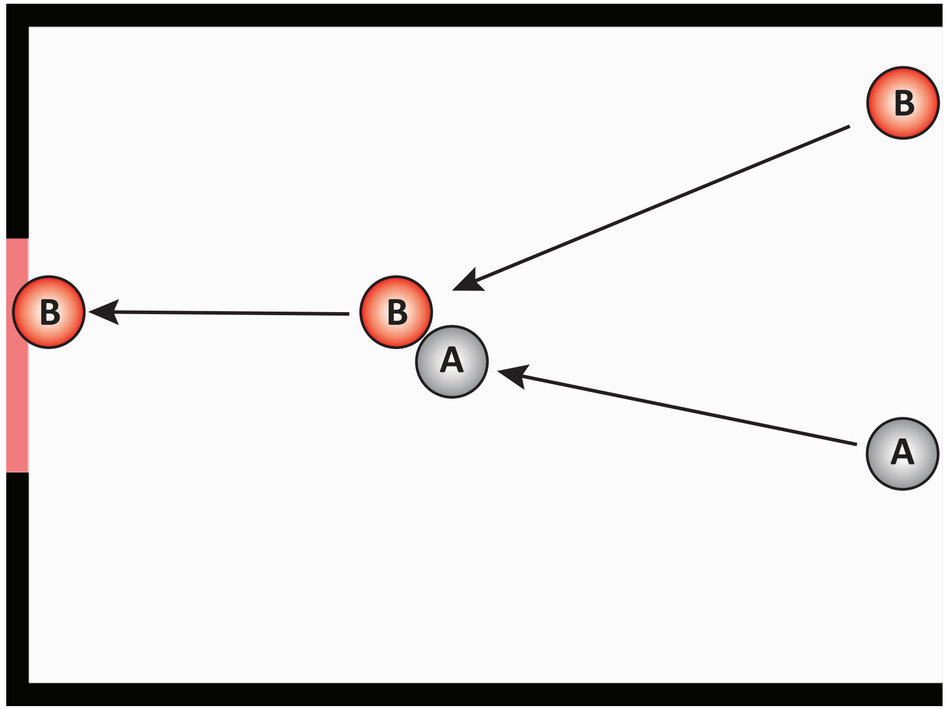
What would have happened?



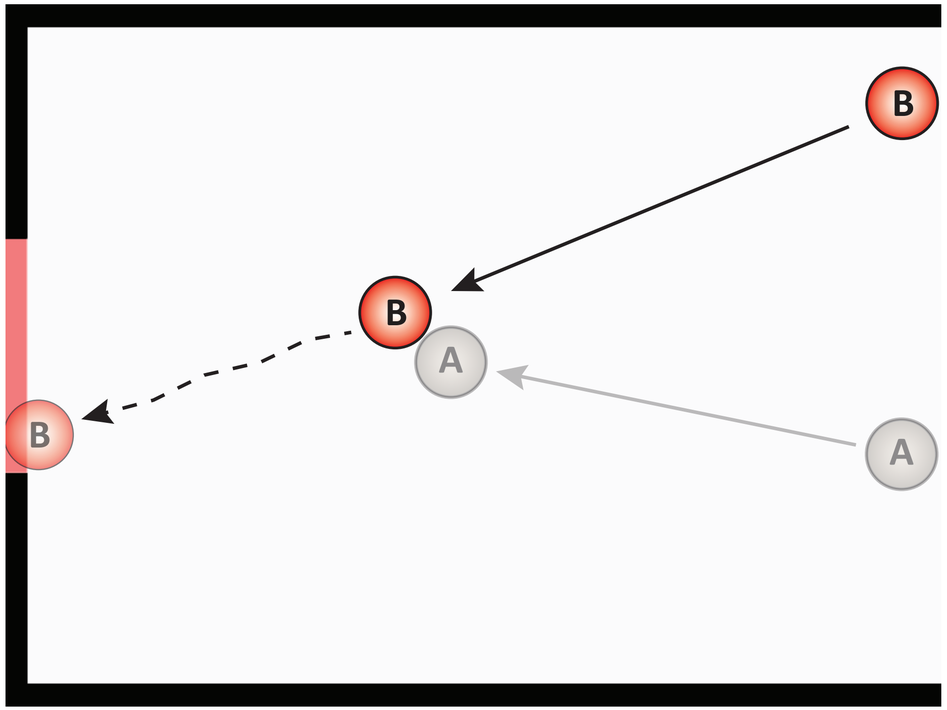
Counterfactual situation

B would have missed the gate ✓

What happened?



What would have happened?



Actual situation

B went through the gate

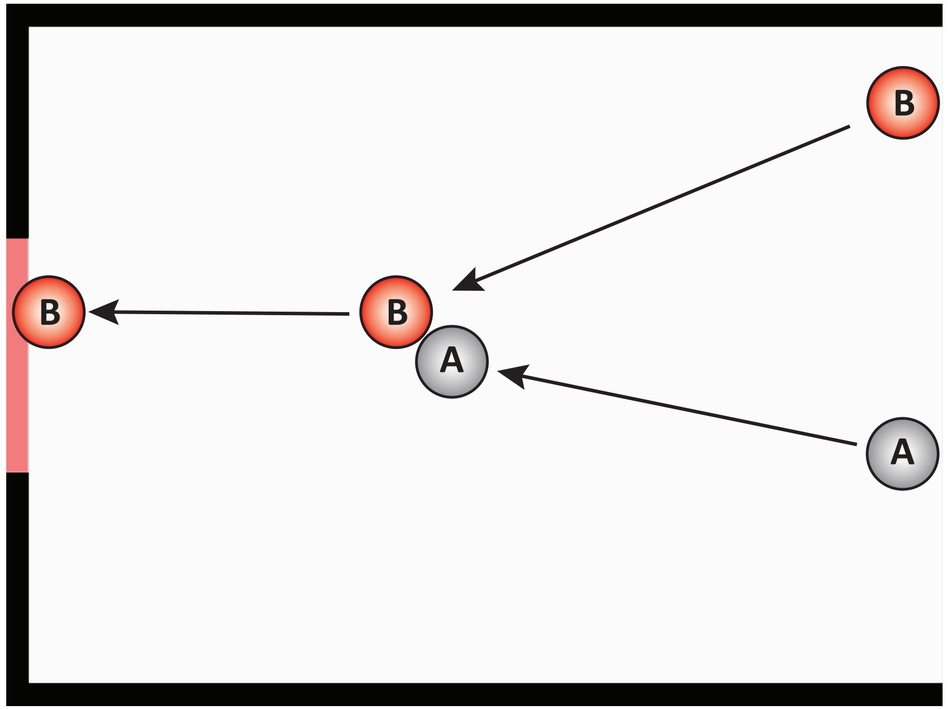
≠

Counterfactual situation

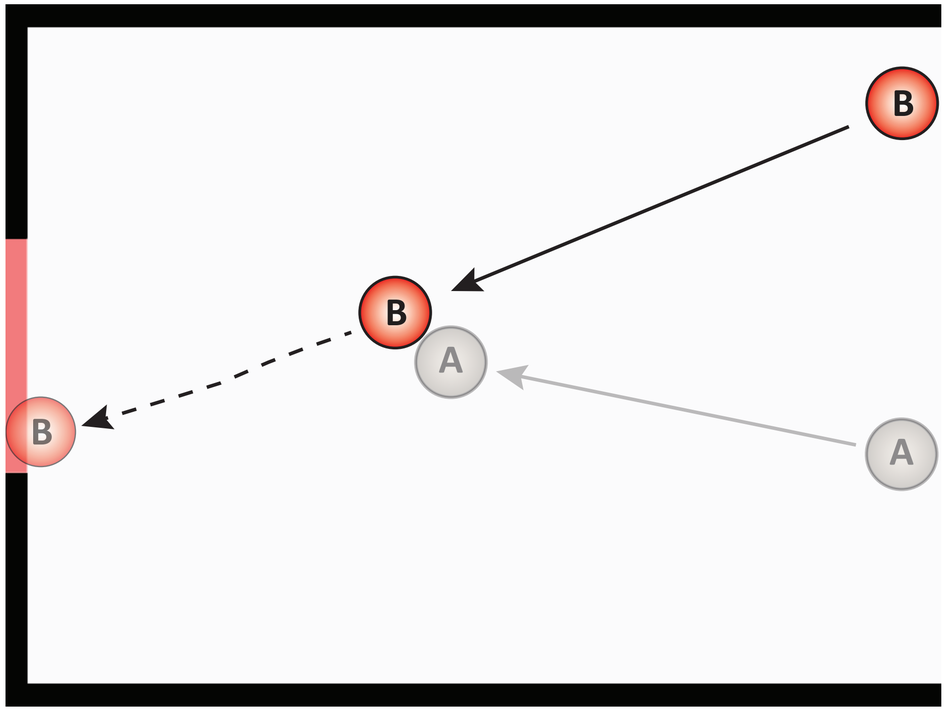
B would have missed the gate ✓

B would have missed the gate ✗

What happened?



What would have happened?



Actual situation

B went through the gate

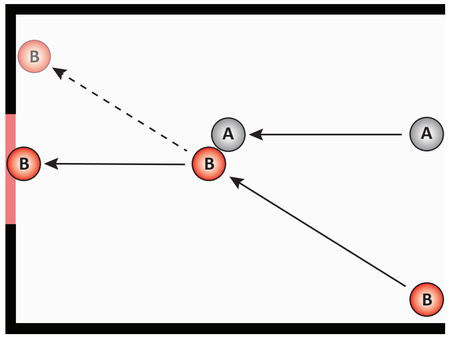
≠

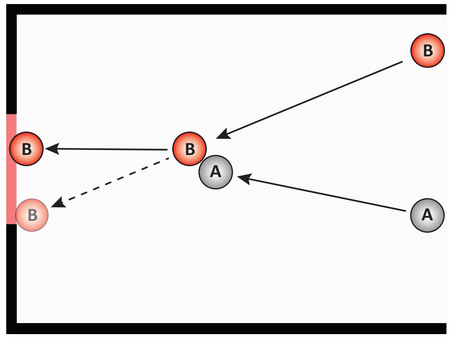
Counterfactual situation

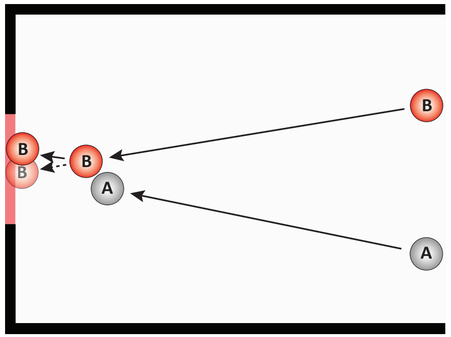
B would have missed the gate ✓

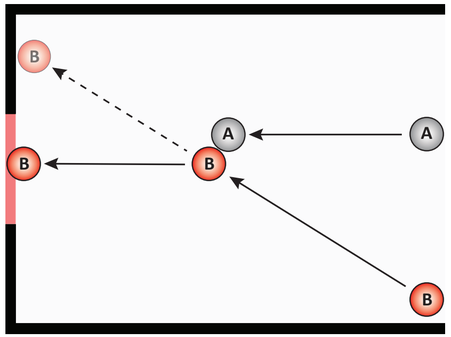
B would have missed the gate ✗

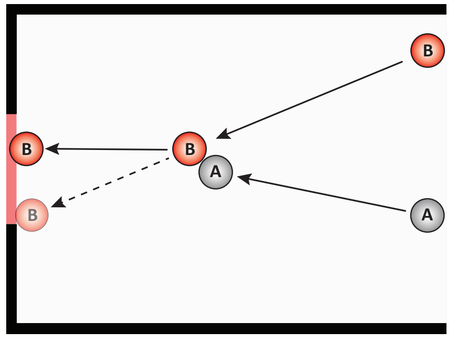
B would have missed the gate ✗

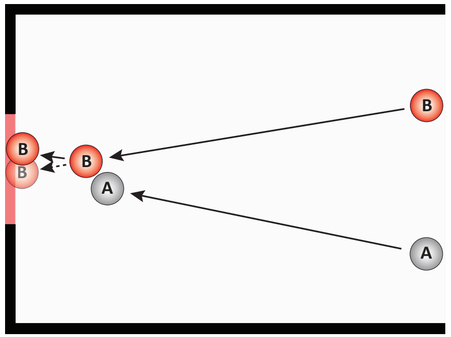

B
A
A
B
B
B


B
B
B
A
B
A


B
B
B
B
A
A



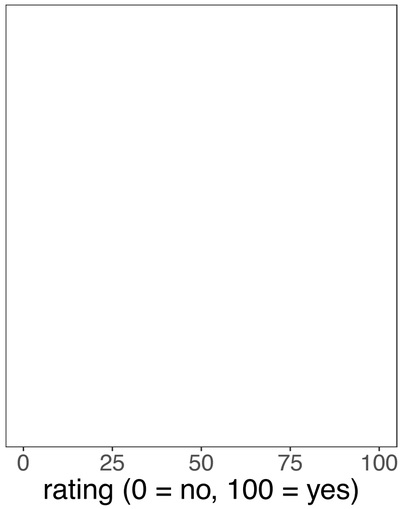


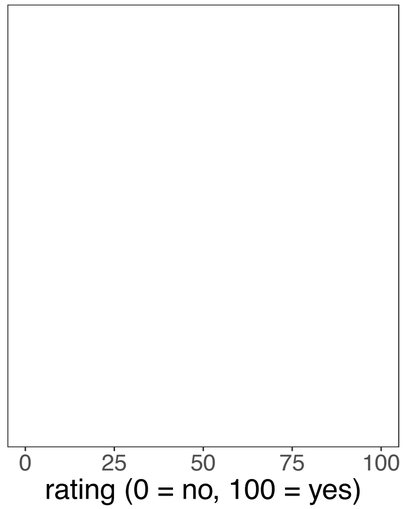


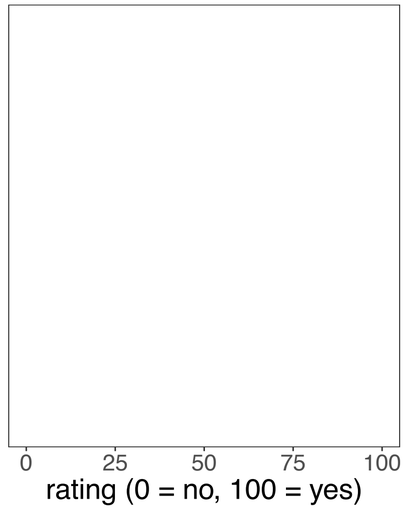
Counterfactual simulation model

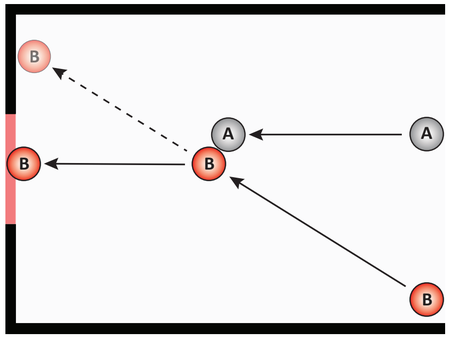
Did A cause B to go through the gate?

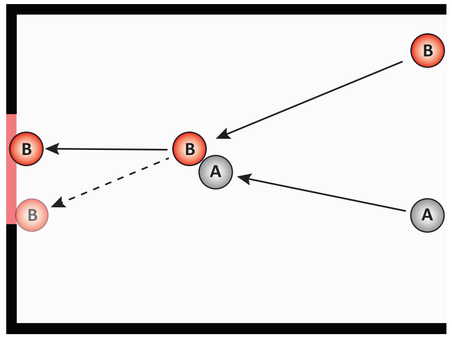
Would B have missed the gate?

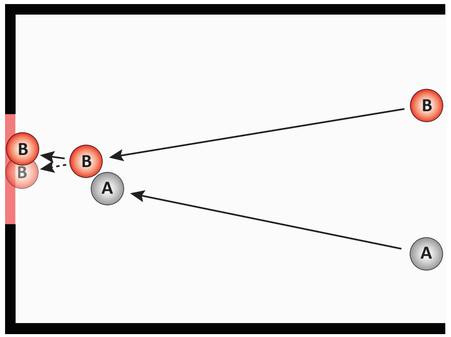








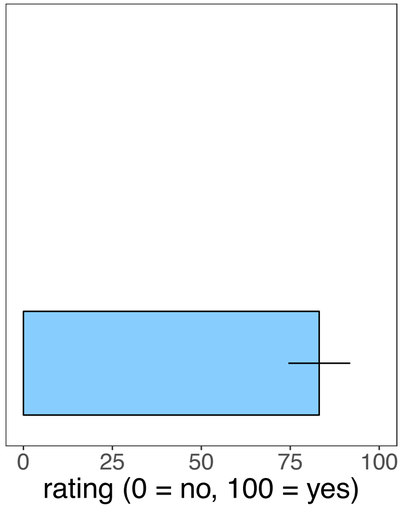


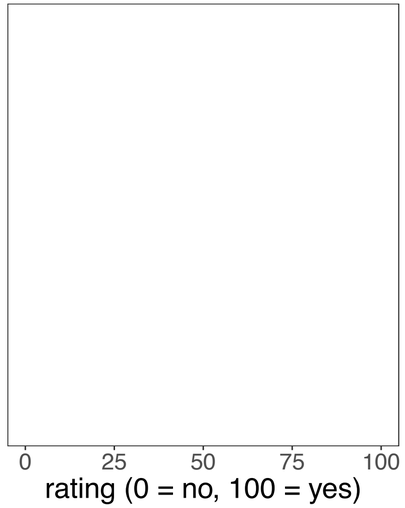


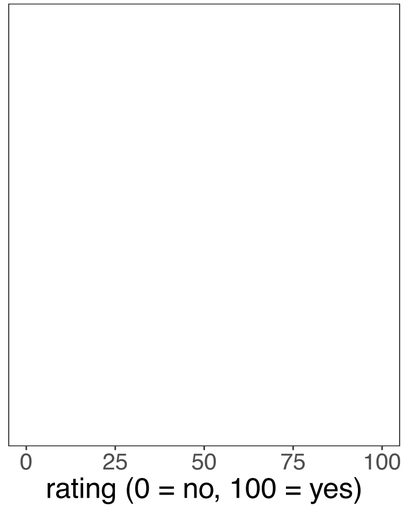
Counterfactual simulation model

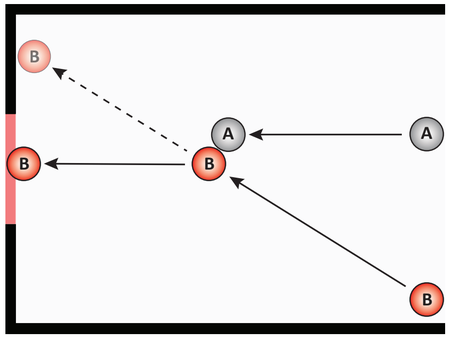
Did A cause B to go through the gate?

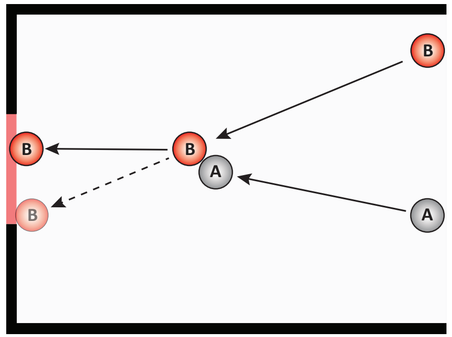
Would B have missed the gate?

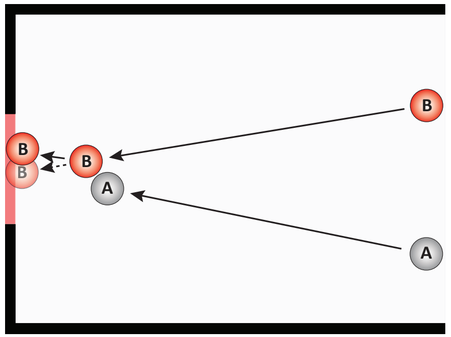








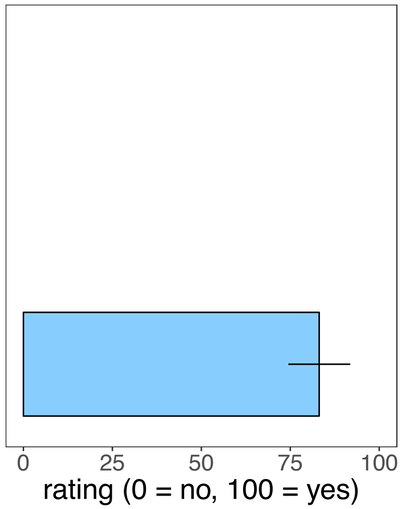


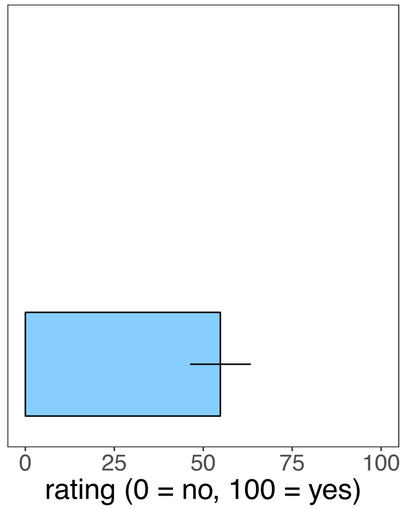


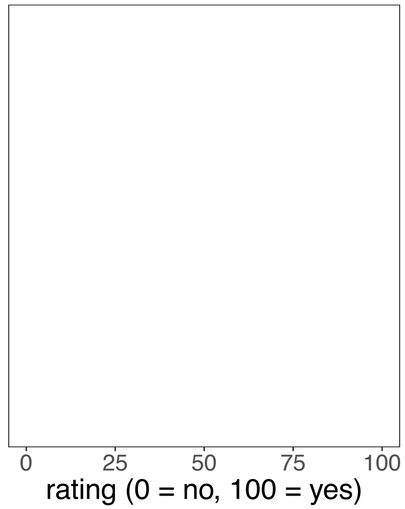
Counterfactual simulation model

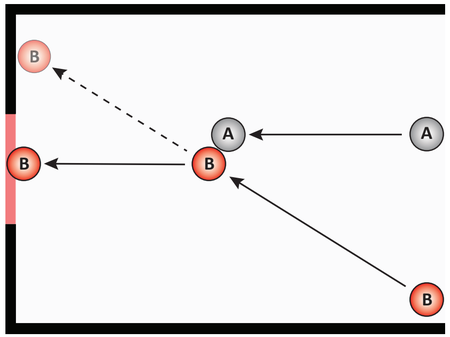
Did A cause B to go through the gate?

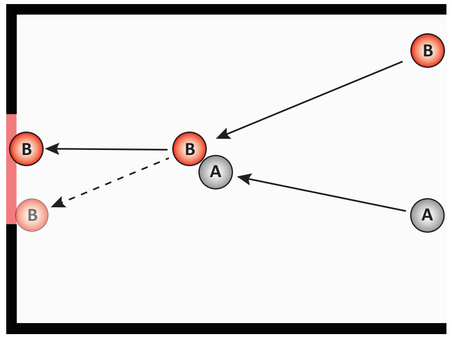
Would B have missed the gate?

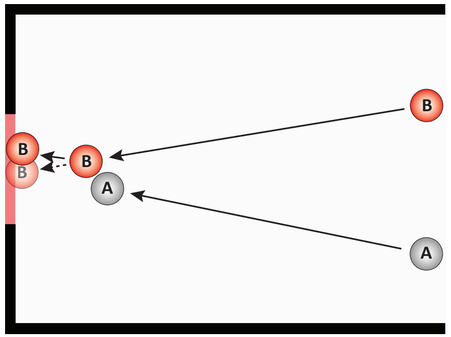








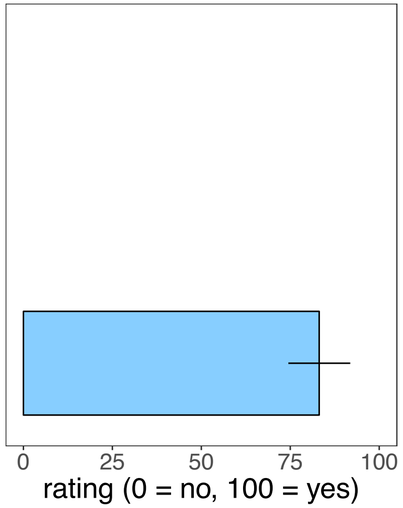


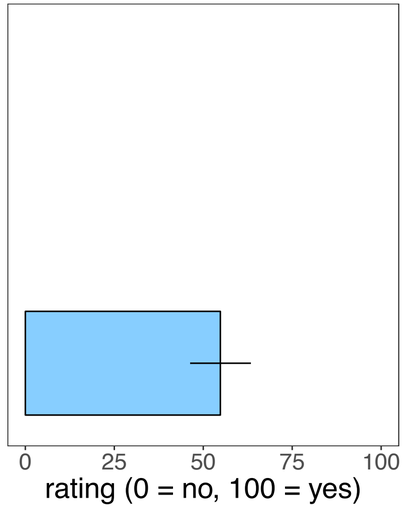


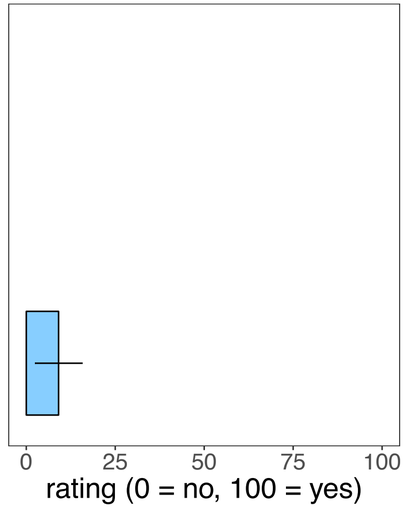
Counterfactual simulation model

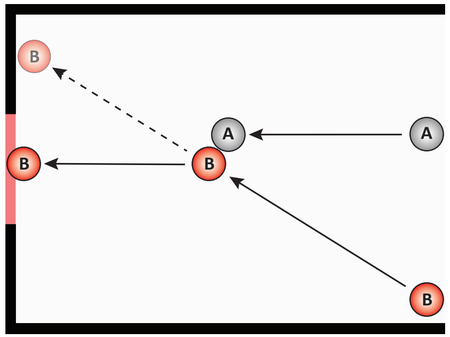
Did A cause B to go through the gate?

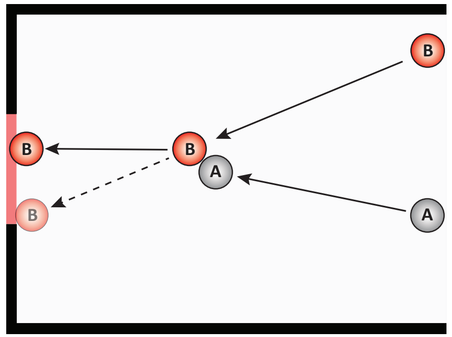
Would B have missed the gate?

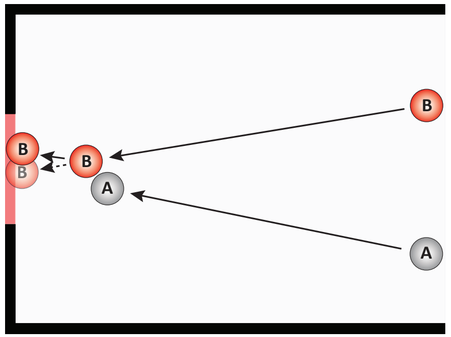








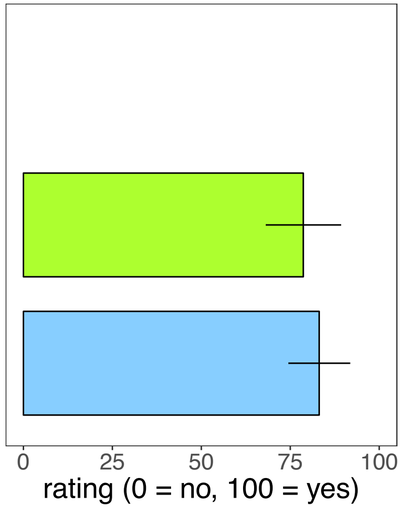


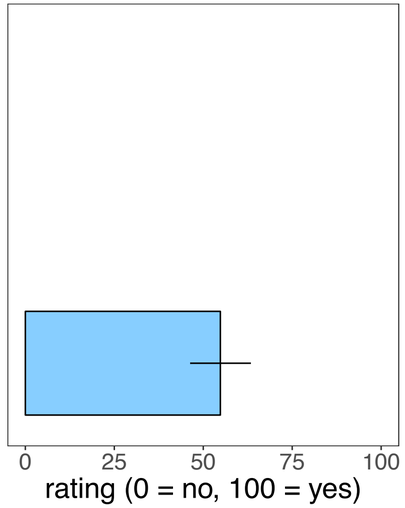


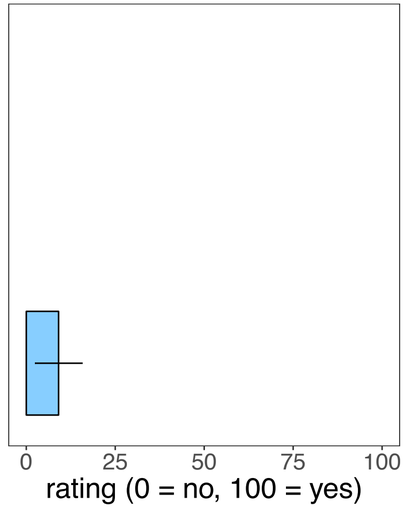
Counterfactual simulation model

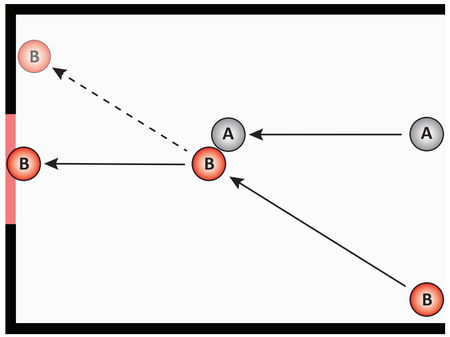
Did A cause B to go through the gate?

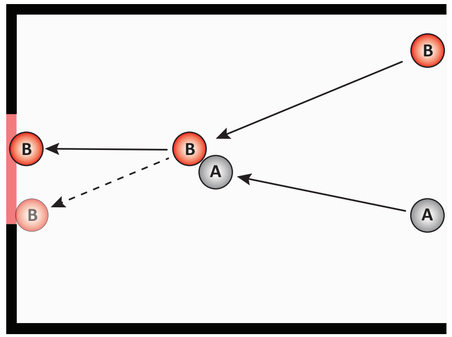
Would B have missed the gate?

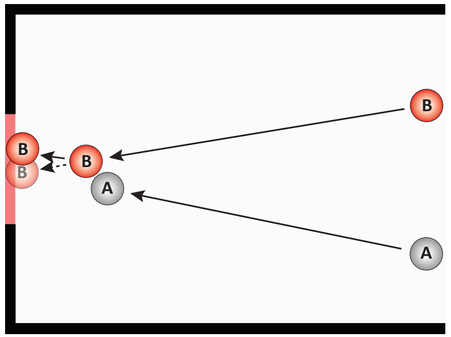








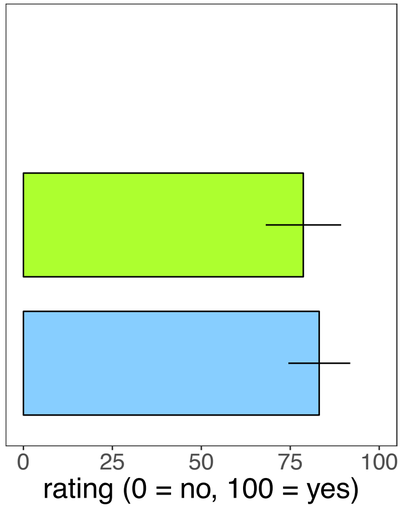


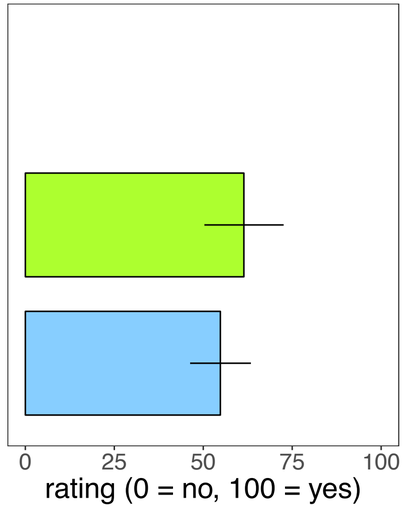


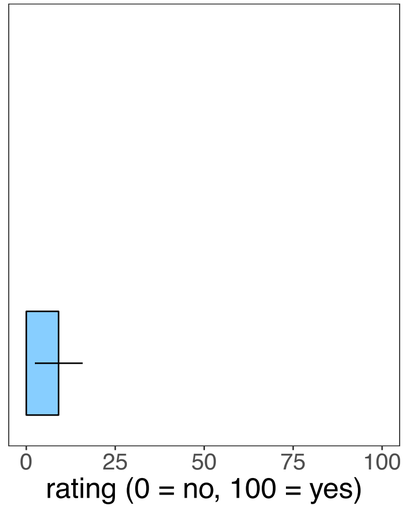
Counterfactual simulation model

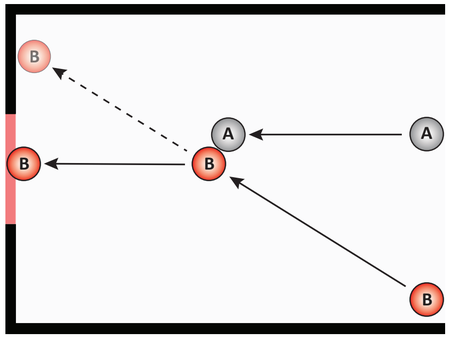
Did A cause B to go through the gate?

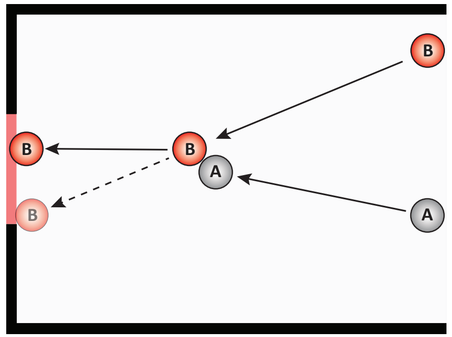
Would B have missed the gate?

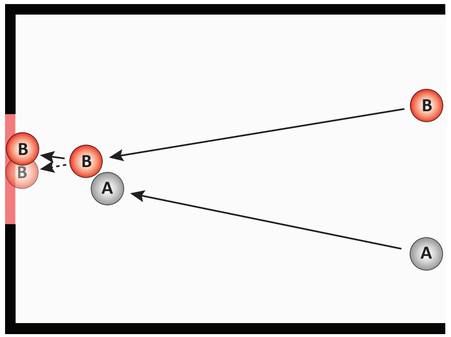








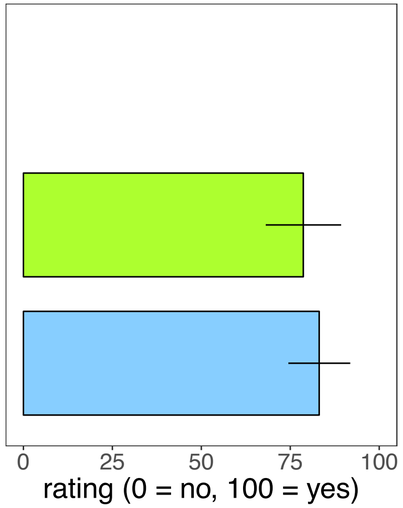


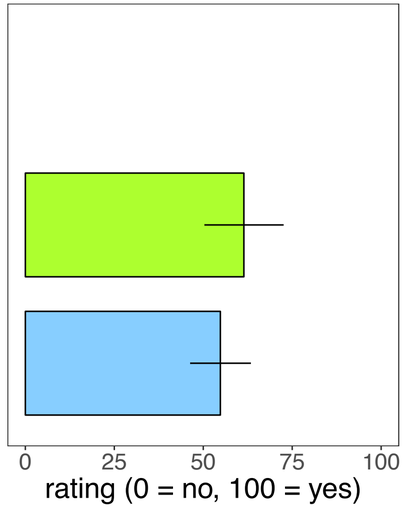


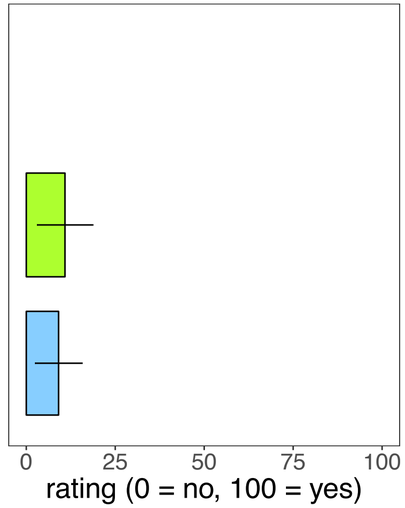
Counterfactual simulation model

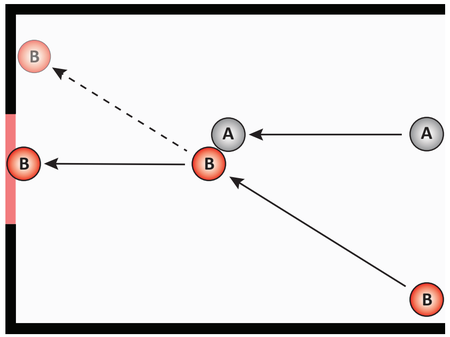
Did A cause B to go through the gate?

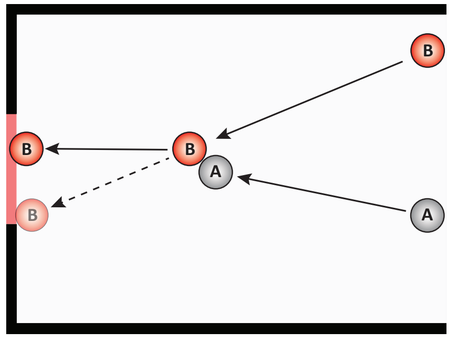
Would B have missed the gate?

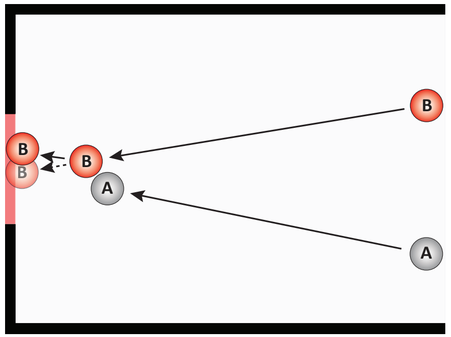








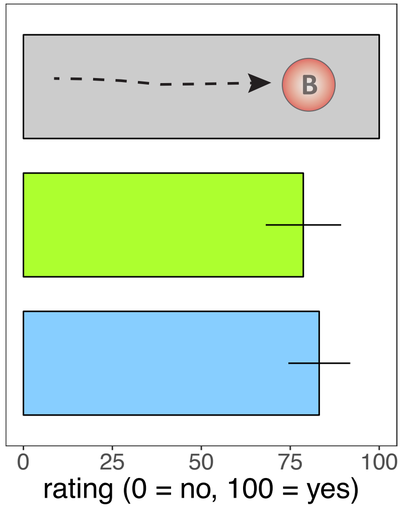


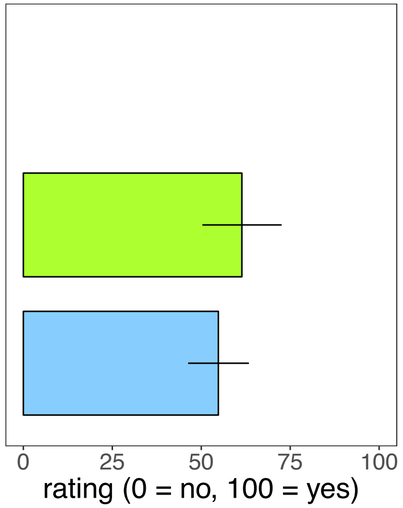


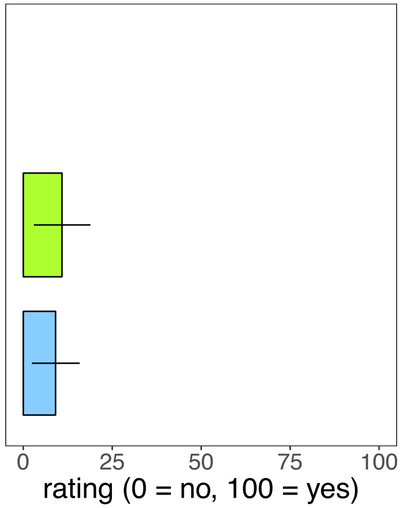
Counterfactual simulation model

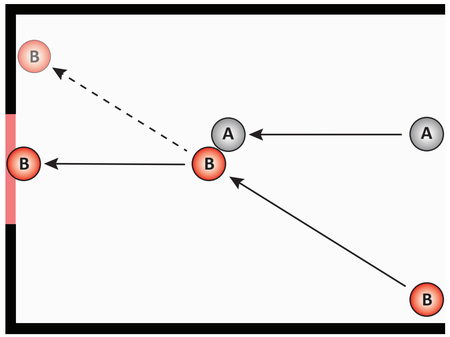
Did A cause B to go through the gate?

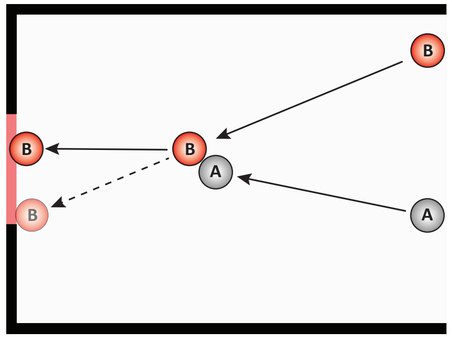
Would B have missed the gate?

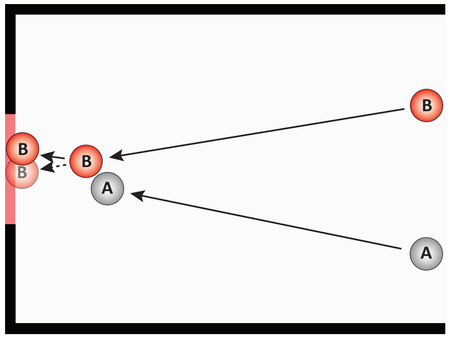








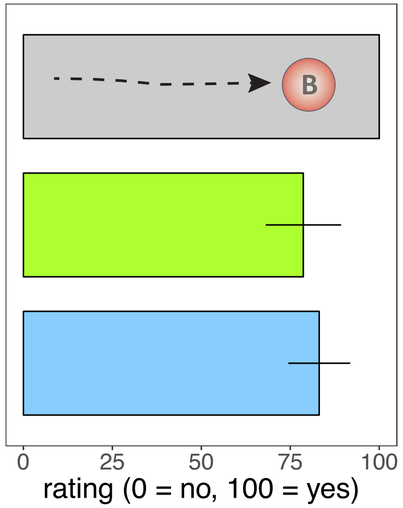


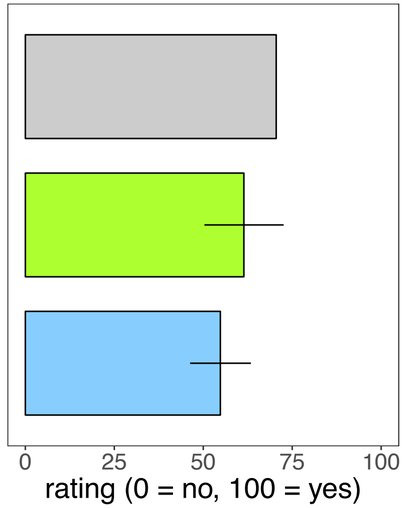


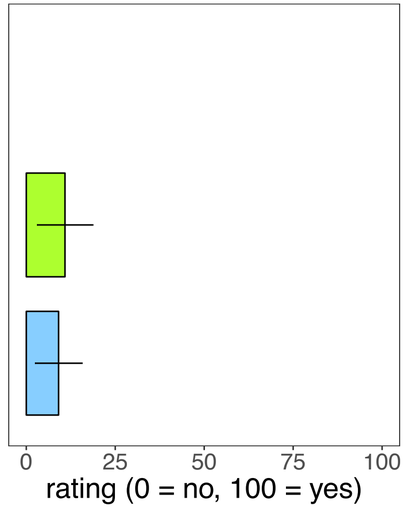
Counterfactual simulation model

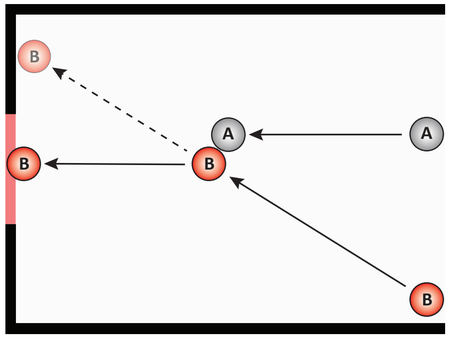
Did A cause B to go through the gate?

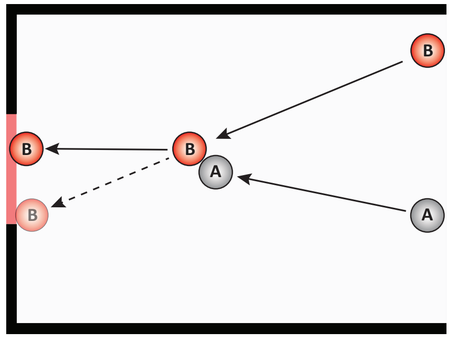
Would B have missed the gate?

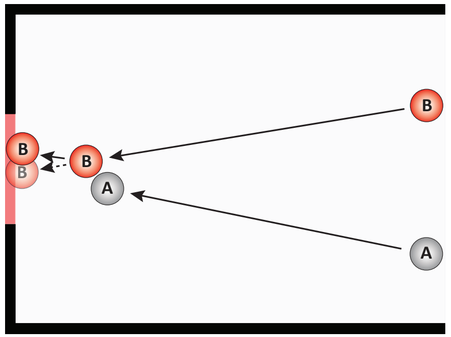








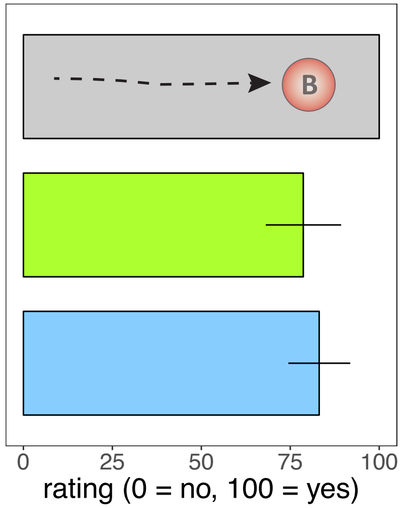


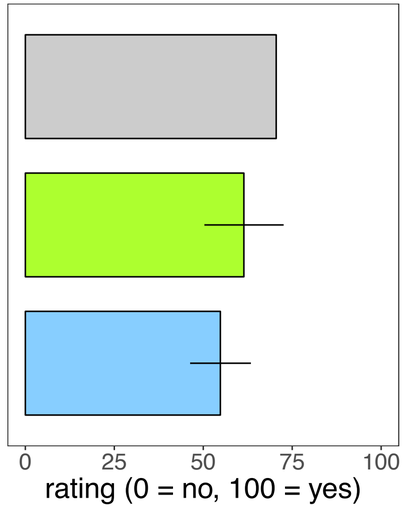


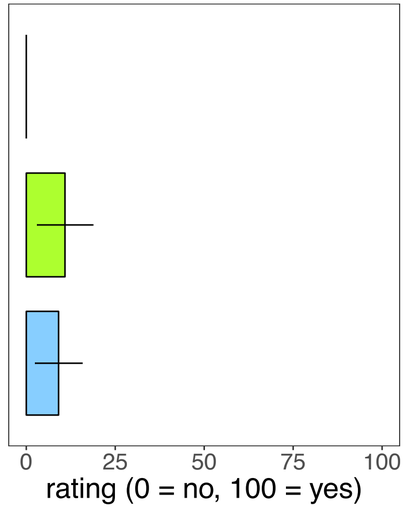
Counterfactual simulation model

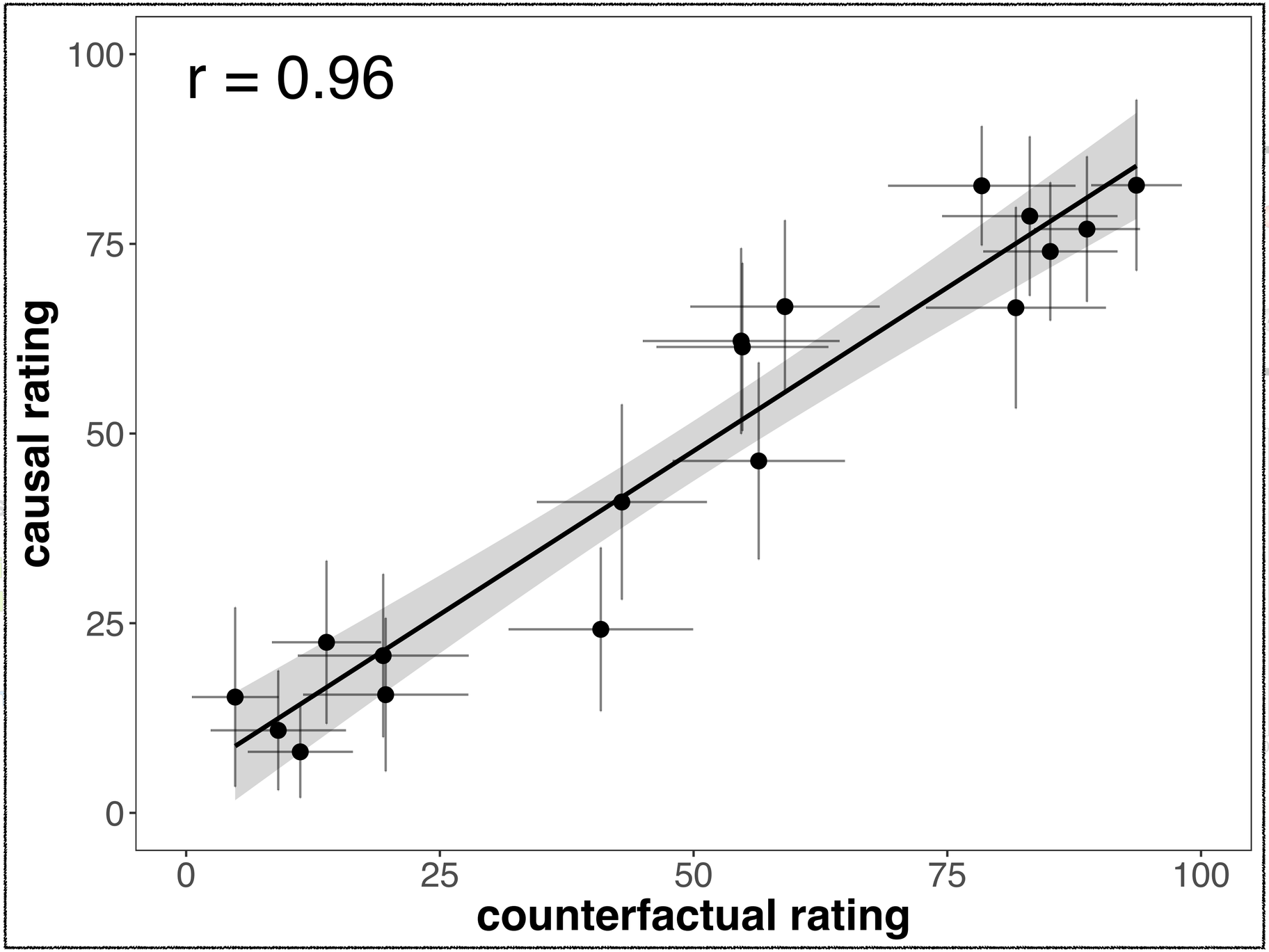
Did A cause B to go through the gate?

Would B have missed the gate?



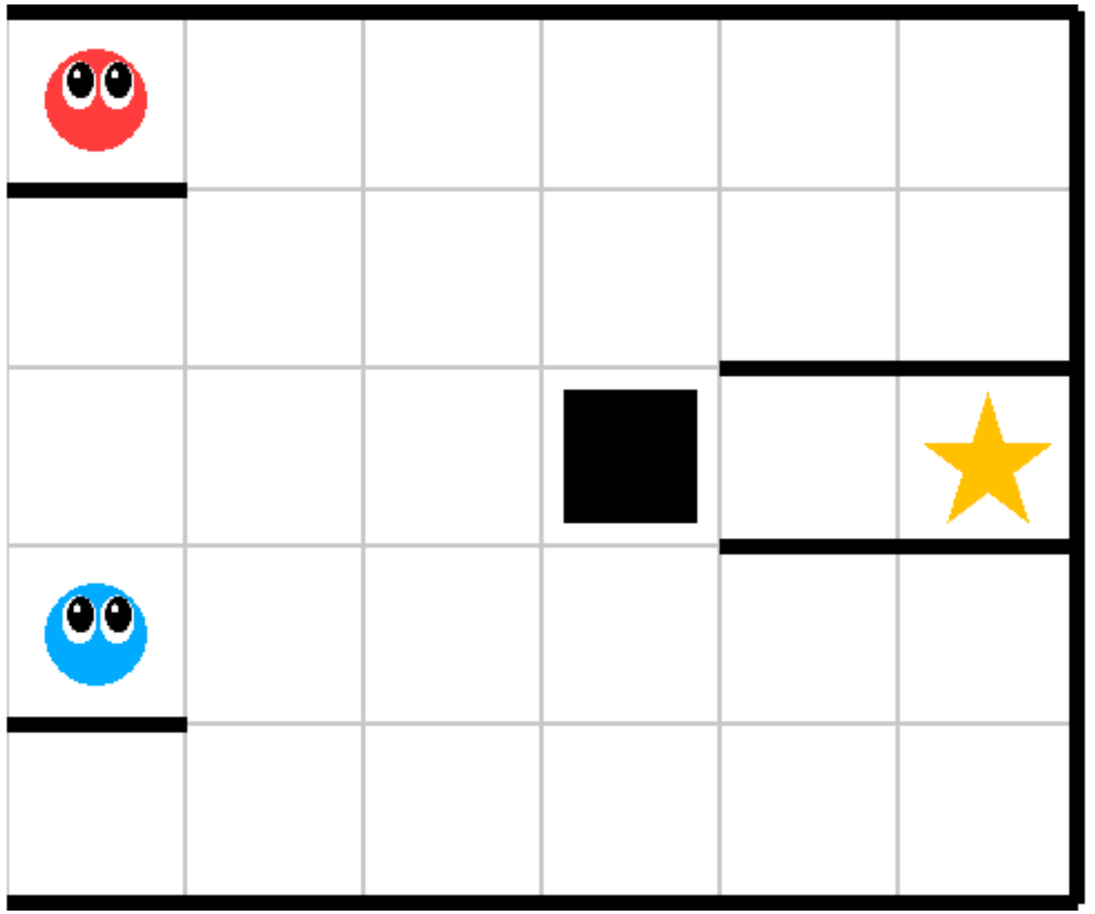





r = 0.96
causal rating
counterfactual rating
100
75
50
25
0
0
25
50
75
100

Deep dive:
Counterfactual simulation for responsibility judgments

Wu et al. "*A computational model of responsibility judgments from counterfactual simulations and intention inferences.*" CogSci, 2023.



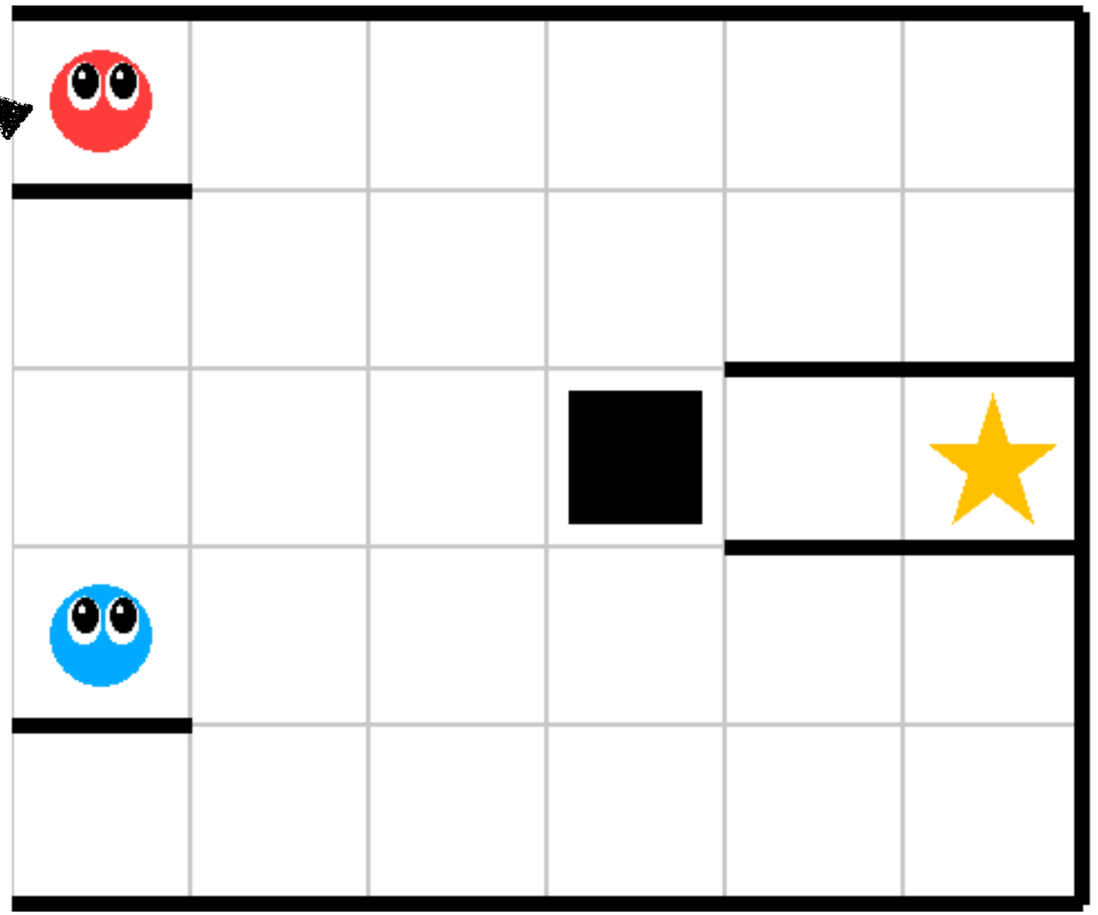
time left:

10

result:

wants to get

to the star

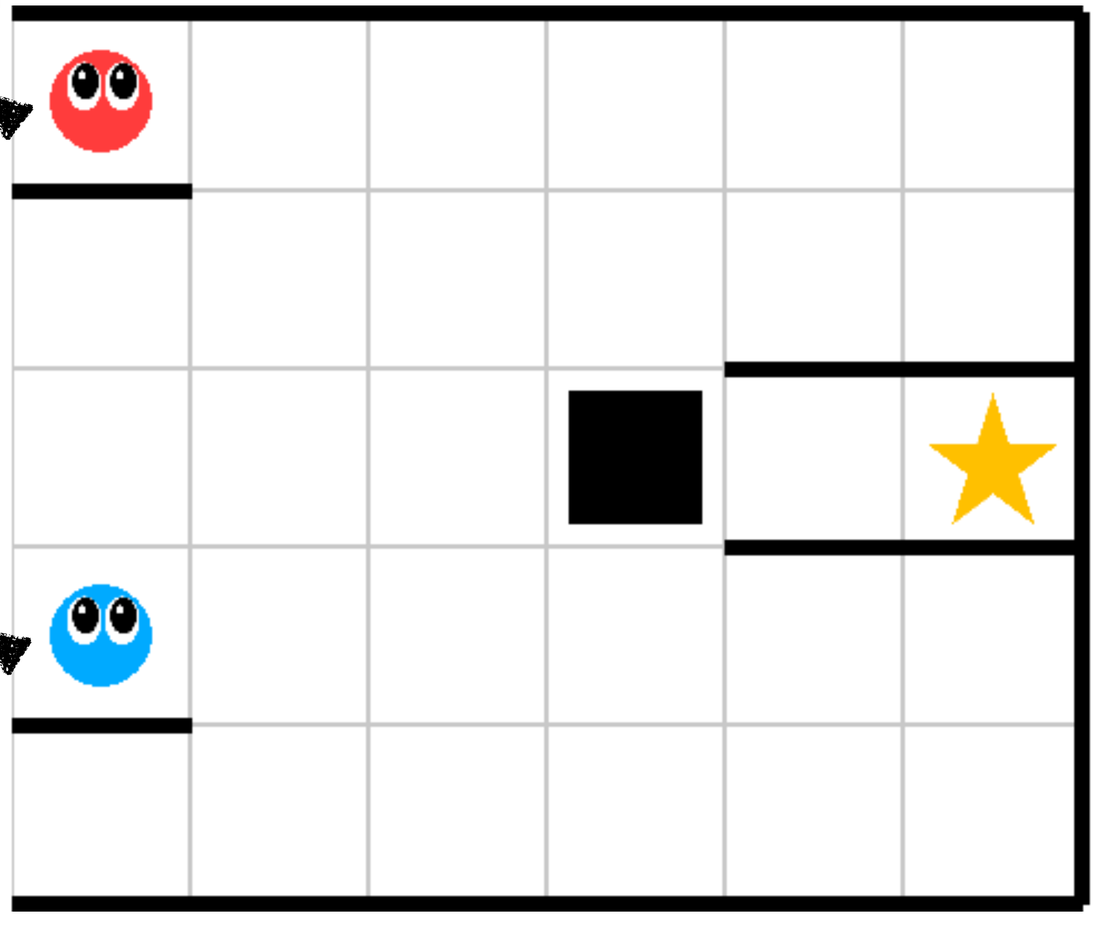


time left:

10

result:

wants to get
to the star



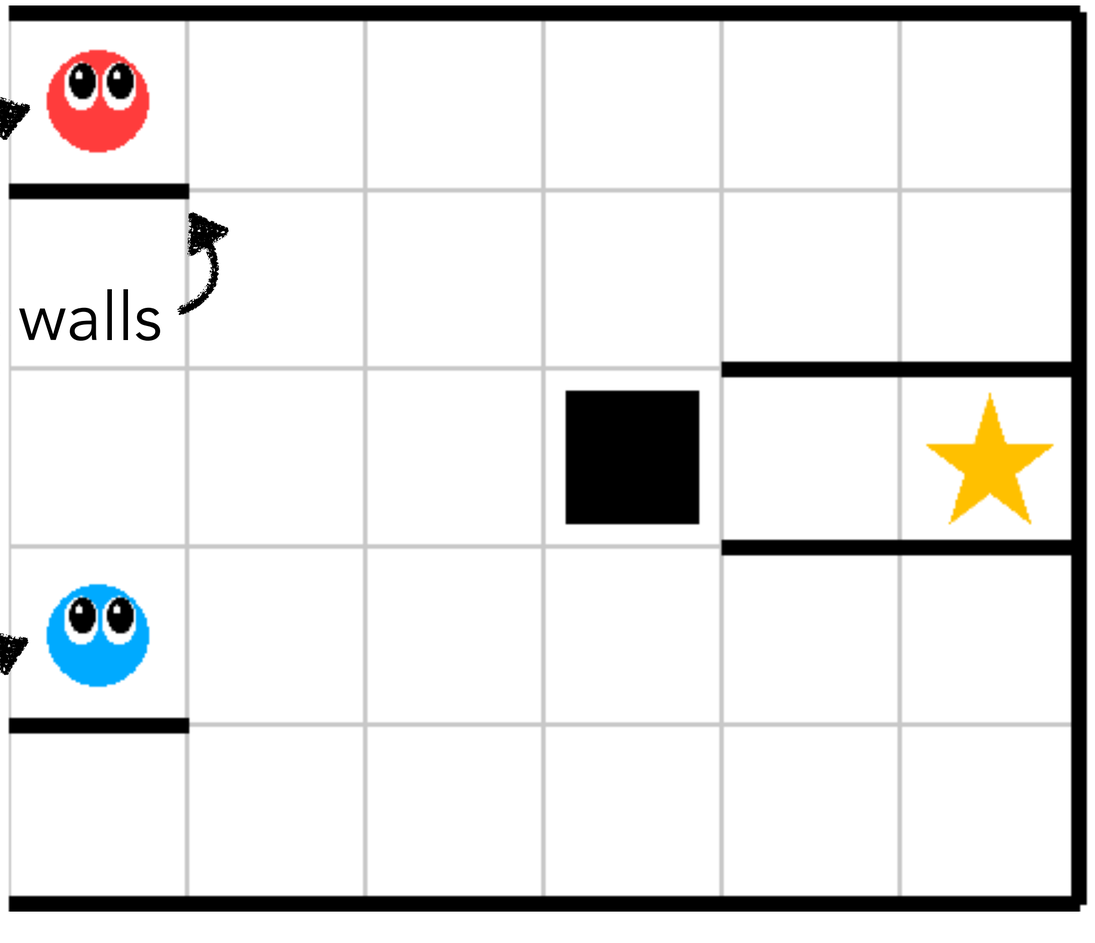
time left:

10

result:

wants to help or
hinder **red**

wants to get
to the star



time left:

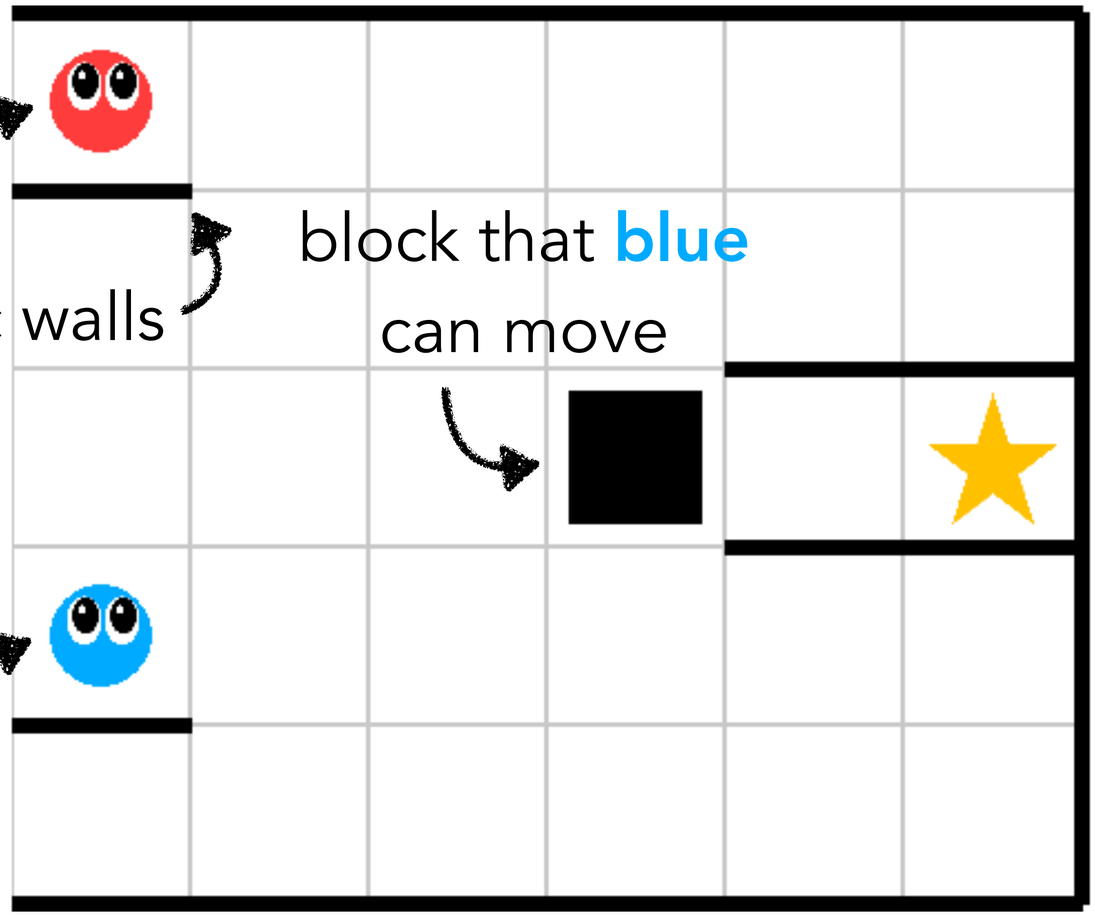
10

static walls

result:

wants to help or
hinder **red**

wants to get
to the star



time left:

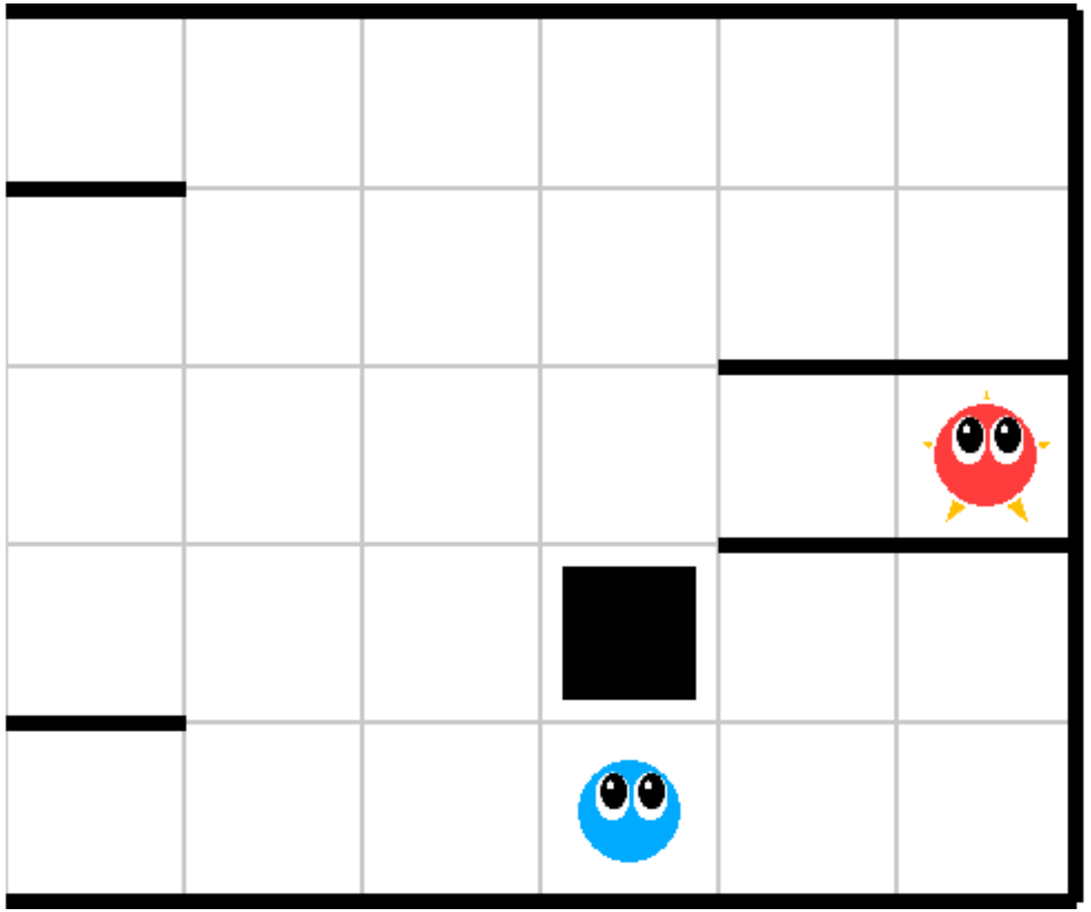
10

block that **blue**
can move

static walls

result:

wants to help or
hinder **red**



time left:

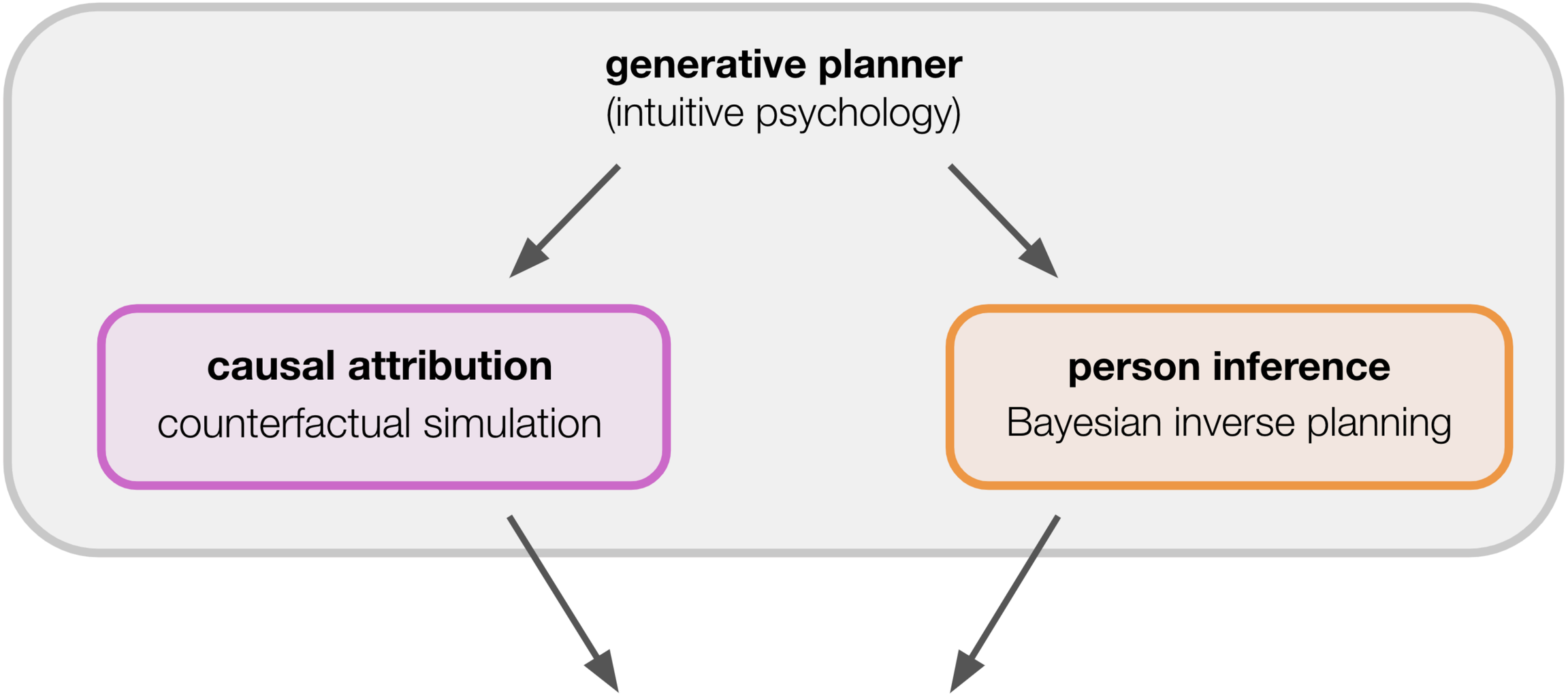
3

result:

SUCCESS

How responsible was the **blue** for the **red's** success?

not at all — very much

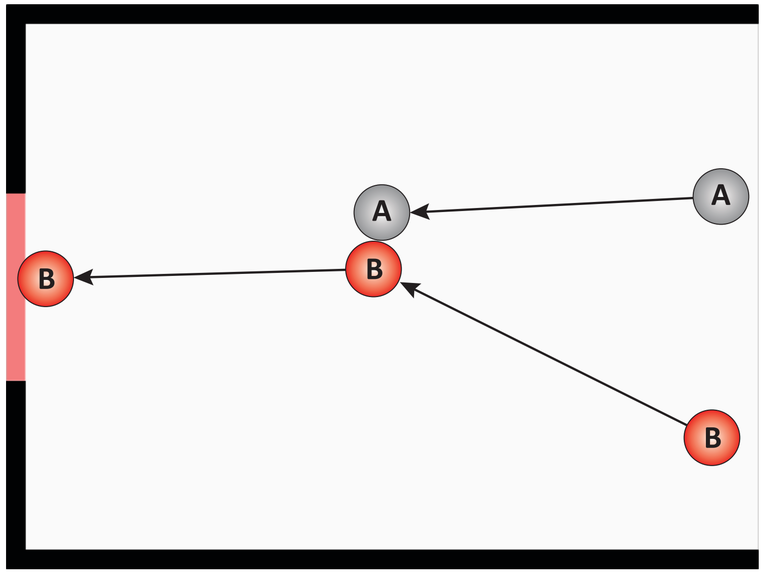

generative planner
(intuitive psychology)
causal attribution
counterfactual simulation
person inference
Bayesian inverse planning
responsibility judgments

causal attribution

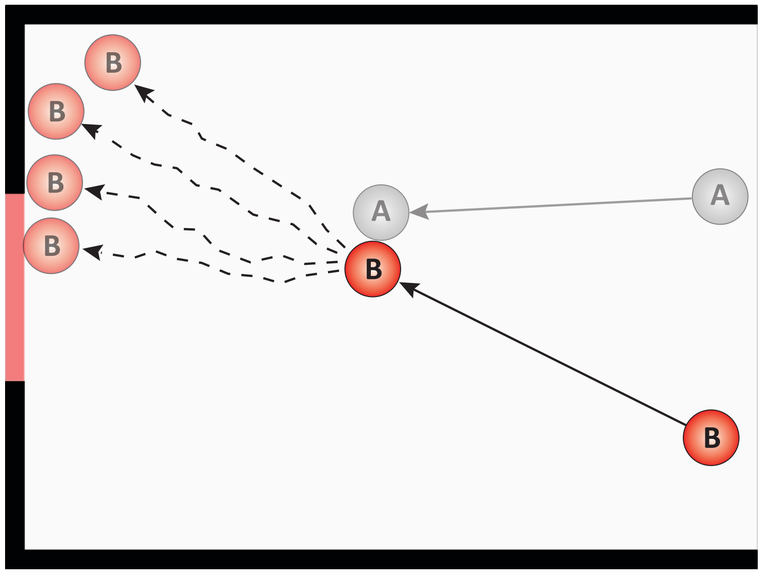
counterfactual simulation

intuitive physics

actual situation



counterfactual simulations



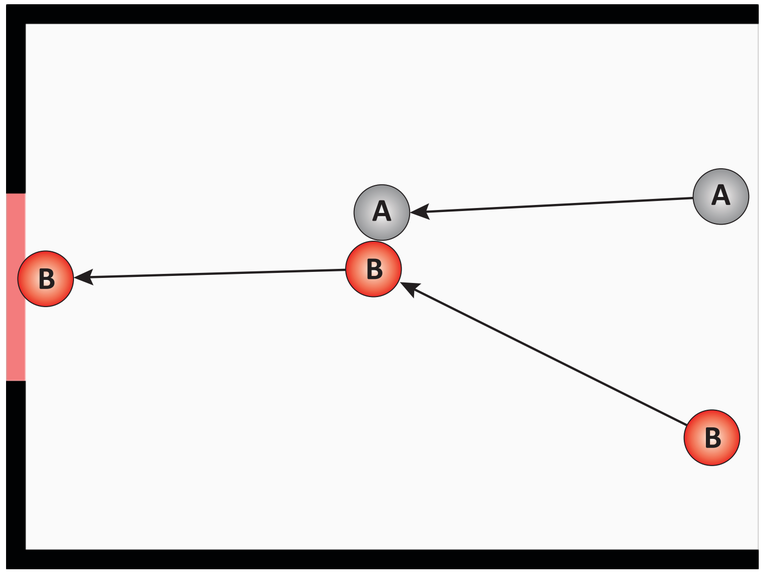
causal attribution

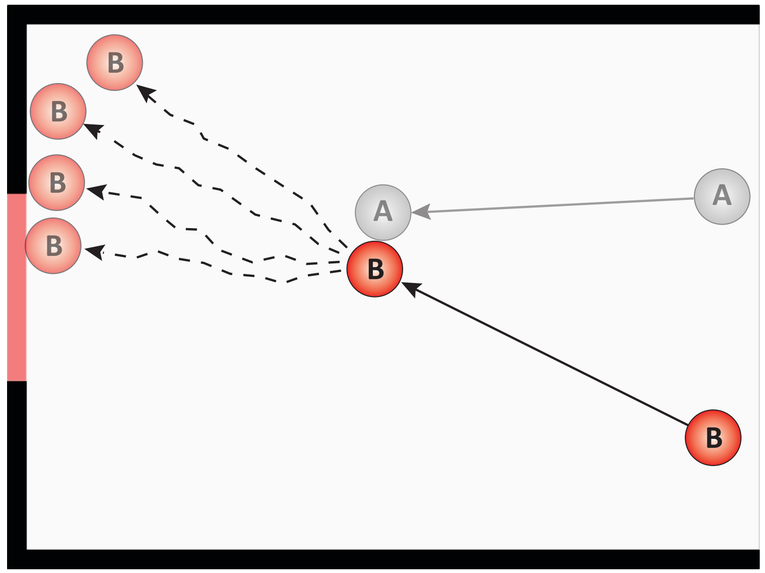
counterfactual simulation

actual situation

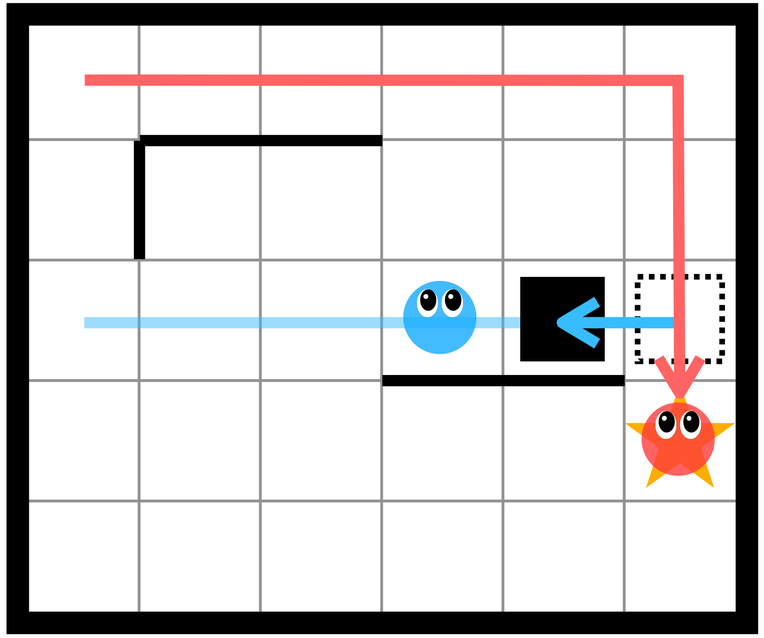
counterfactual simulations

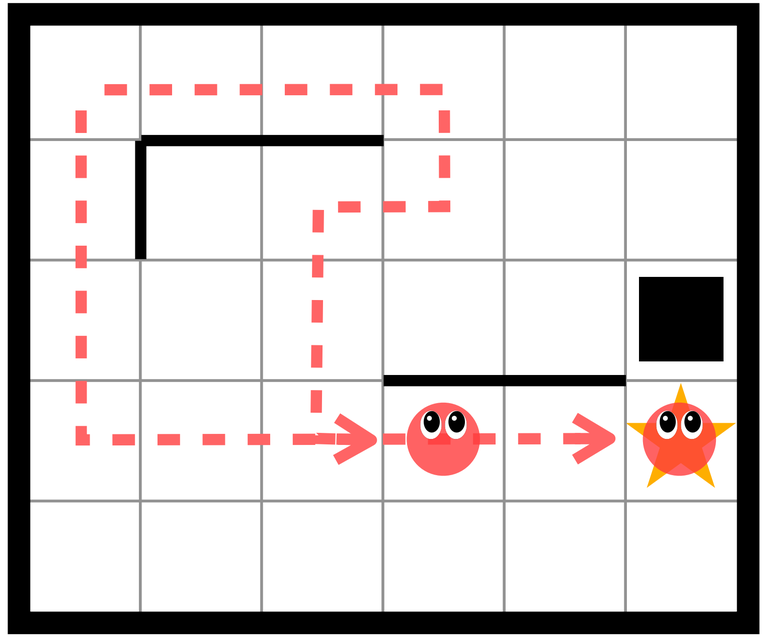
intuitive physics

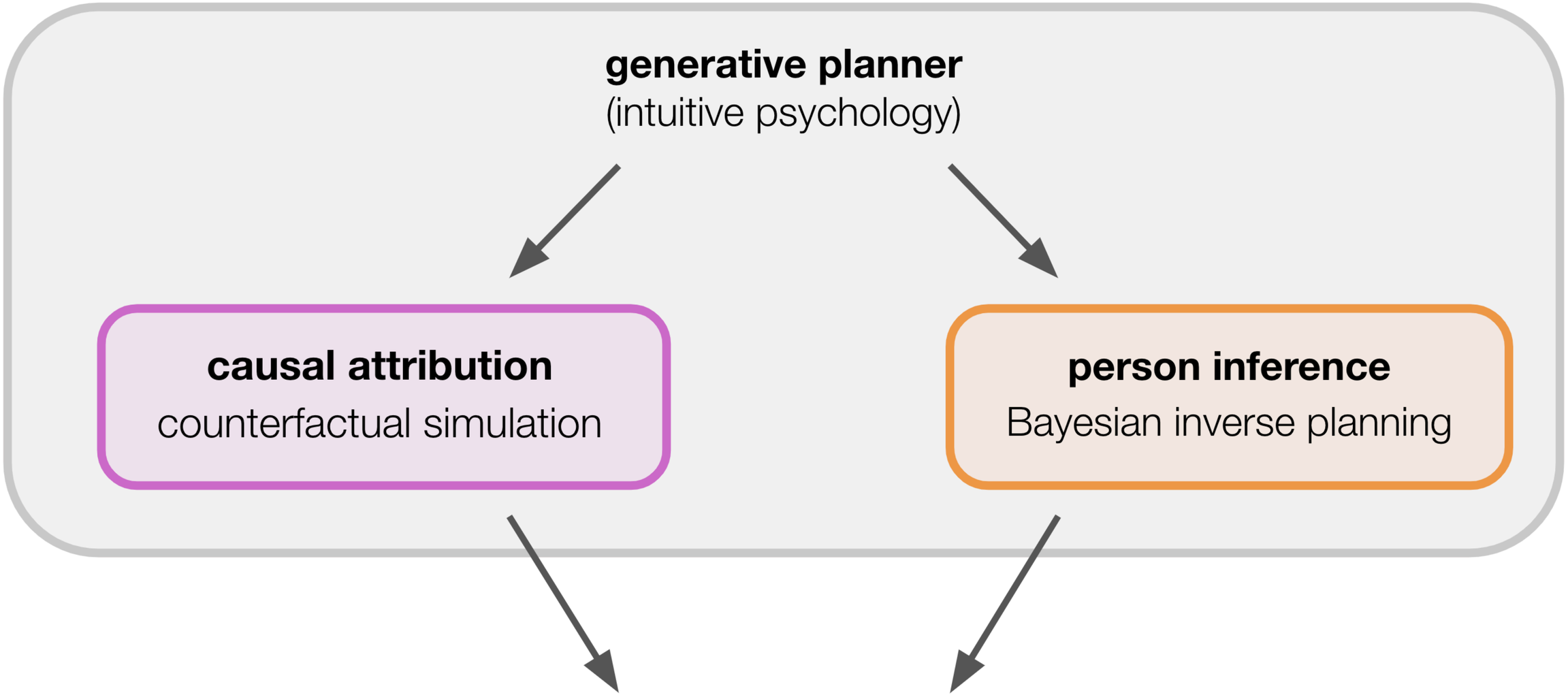




intuitive psychology





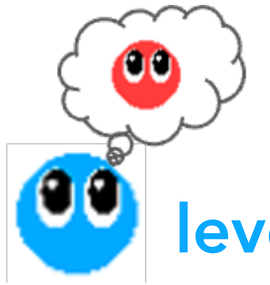

generative planner
(intuitive psychology)
causal attribution
counterfactual simulation
person inference
Bayesian inverse planning
responsibility judgments

person inference

Bayesian inverse planning



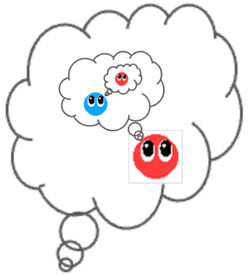
level-0 red plans around obstacles to reach the star



level-1 blue plans to help or hinder a **level-0 red**



level-2 red plans around **level-1 blue** to reach the star

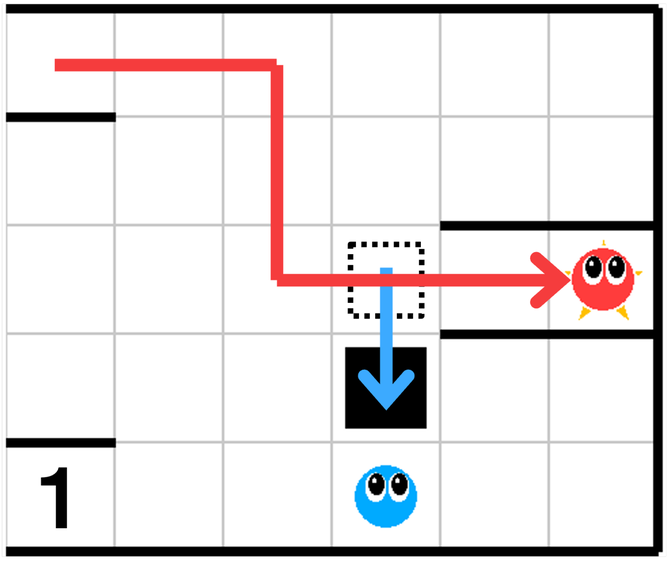




level-3 blue plans to help or deceive a **level-2 red**

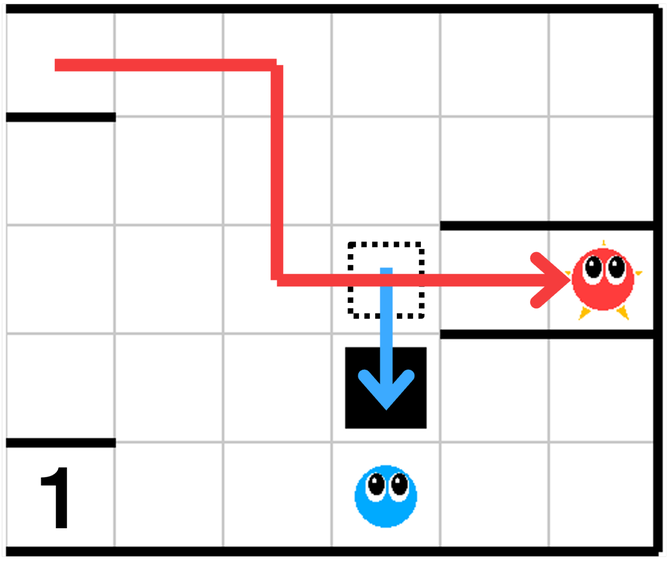
causal attribution
counterfactual simulation

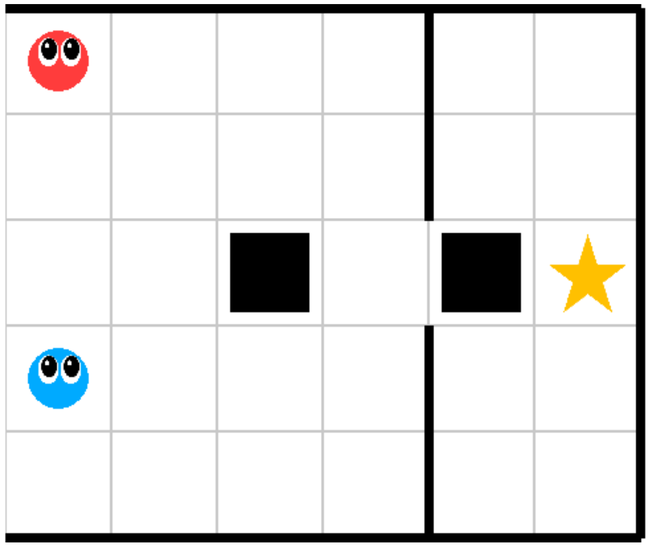
person inference
Bayesian inverse planning



causal attribution
counterfactual simulation

person inference
Bayesian inverse planning





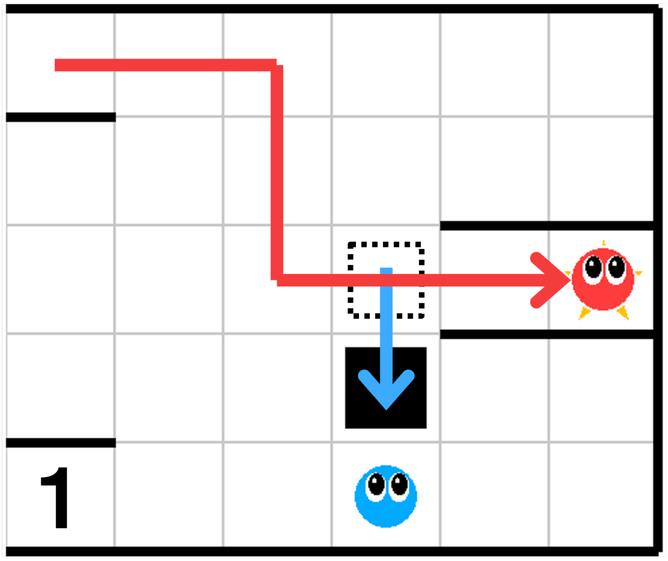
time left:

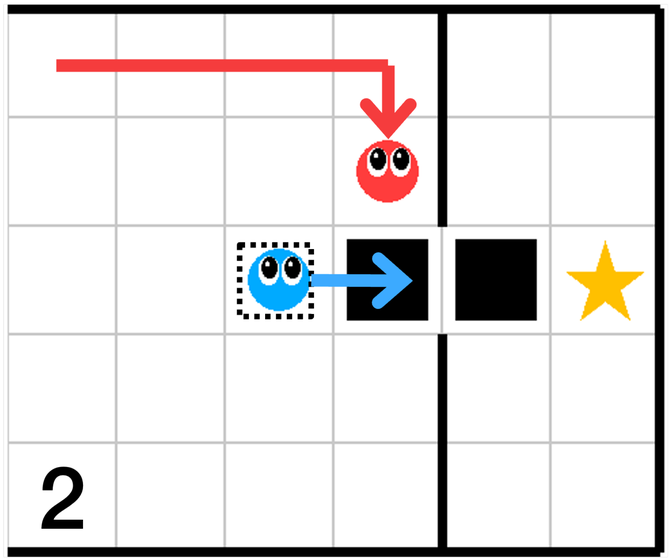
10

result:

causal attribution
counterfactual simulation

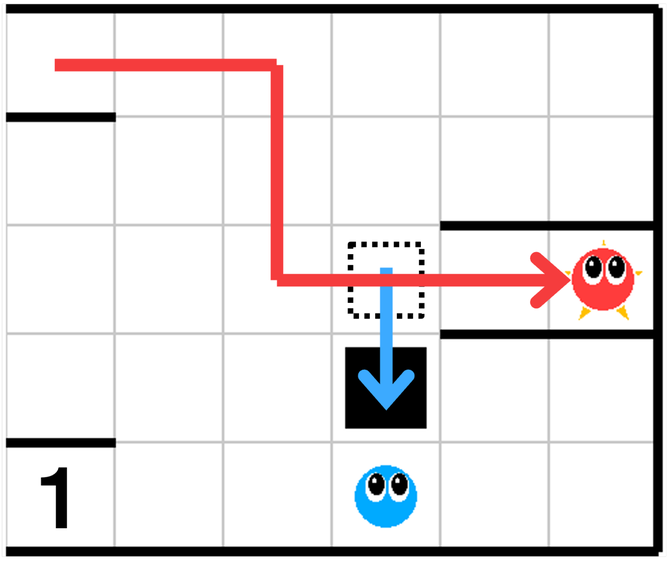
person inference
Bayesian inverse planning

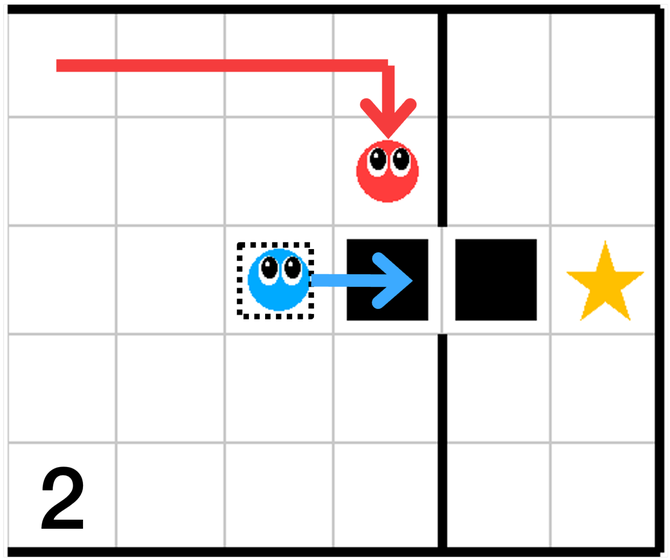


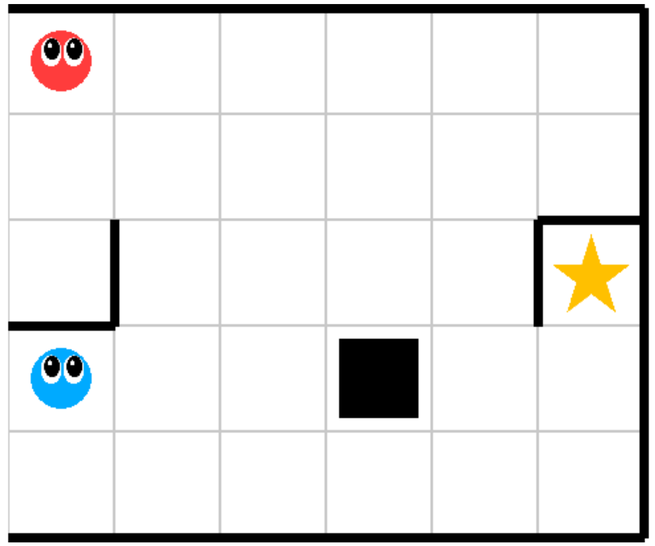


causal attribution
counterfactual simulation

person inference
Bayesian inverse planning





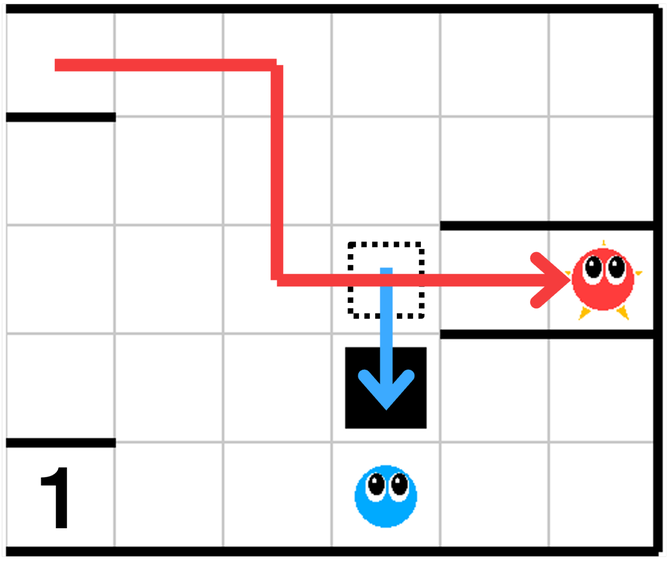


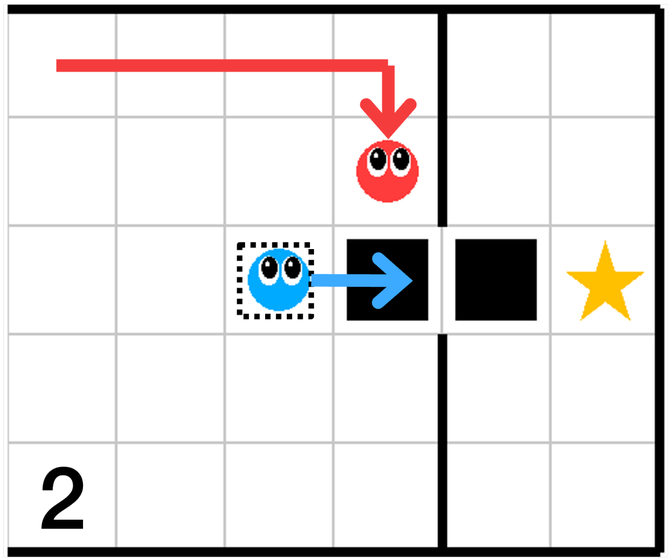
time left:
10

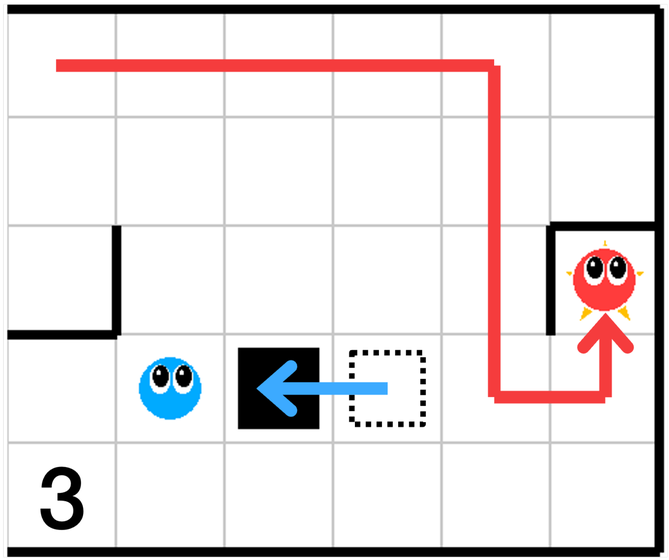
result:

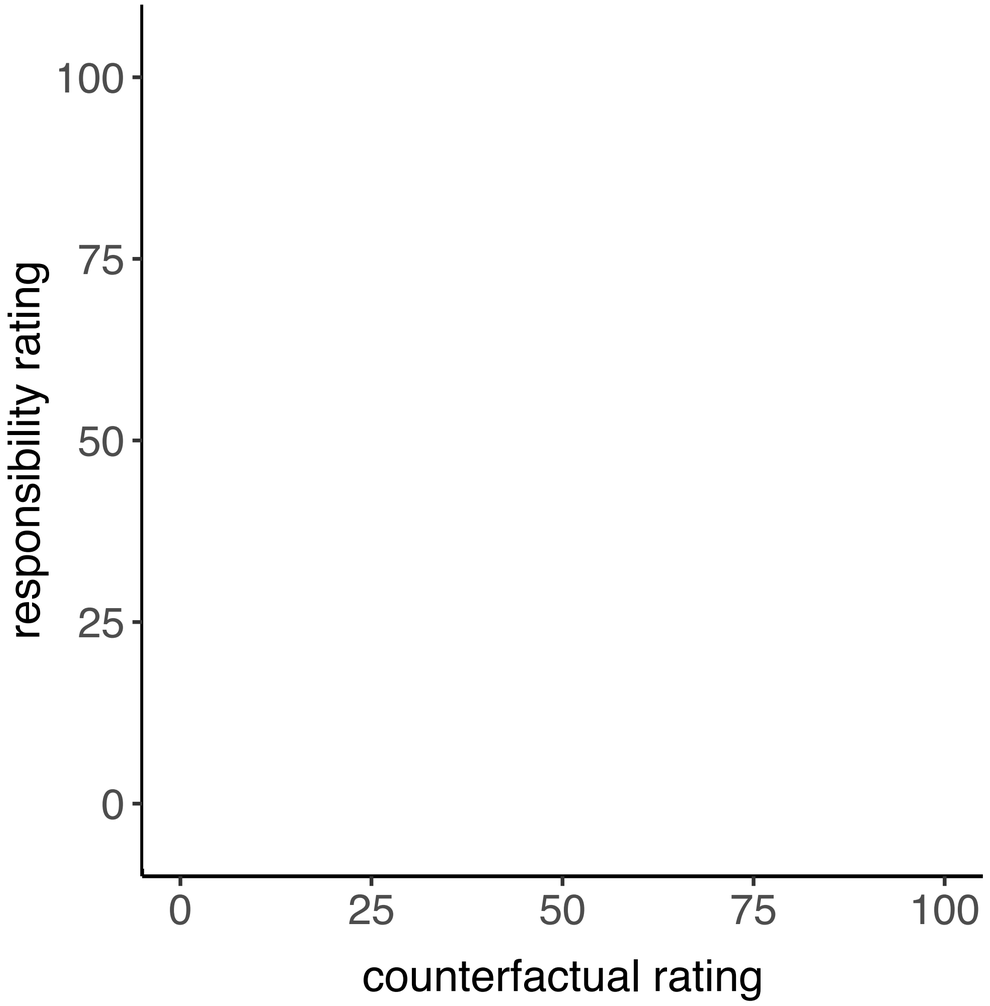
causal attribution
counterfactual simulation

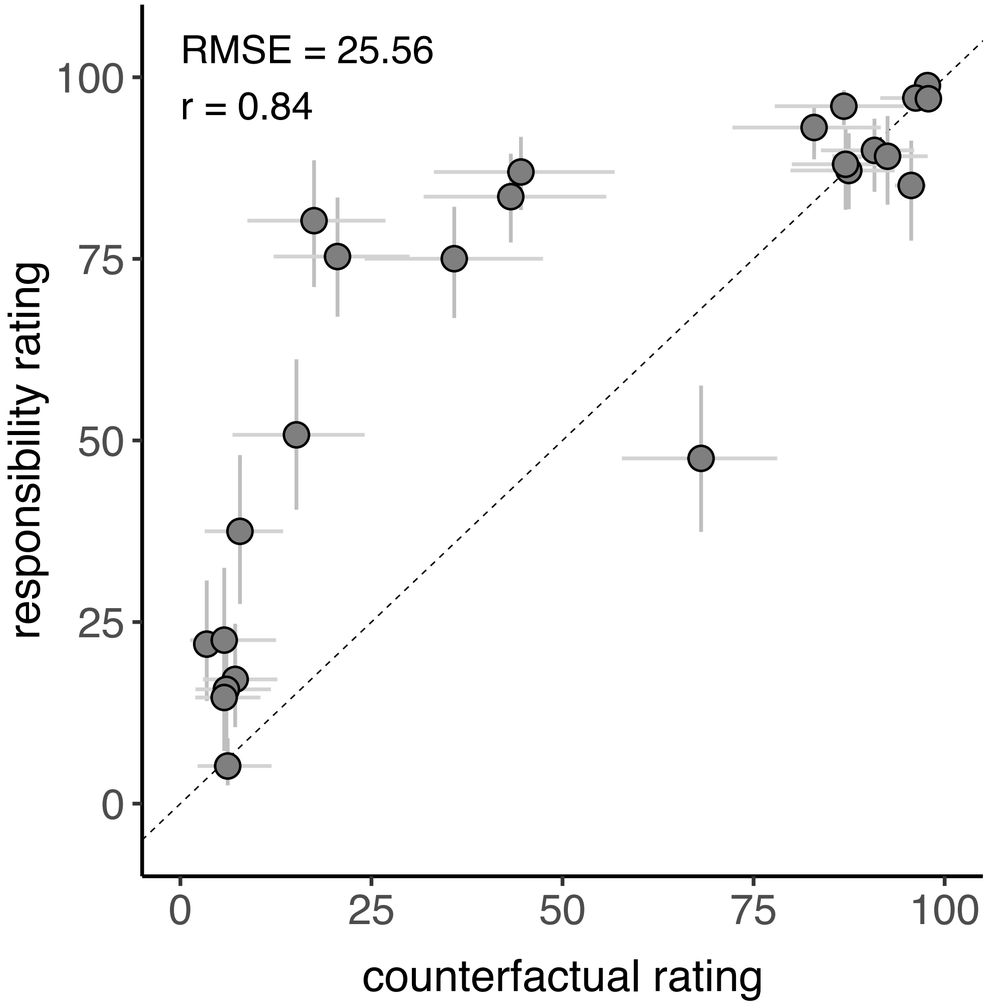
person inference
Bayesian inverse planning

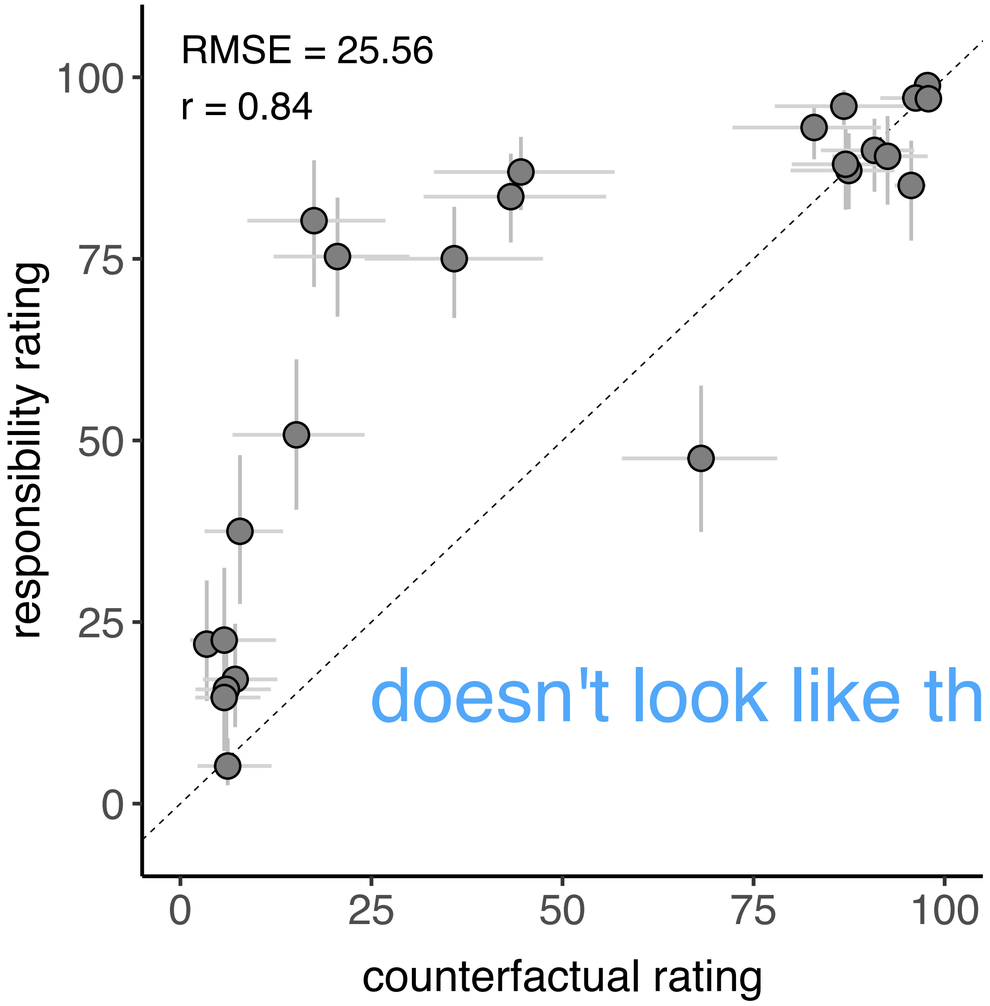




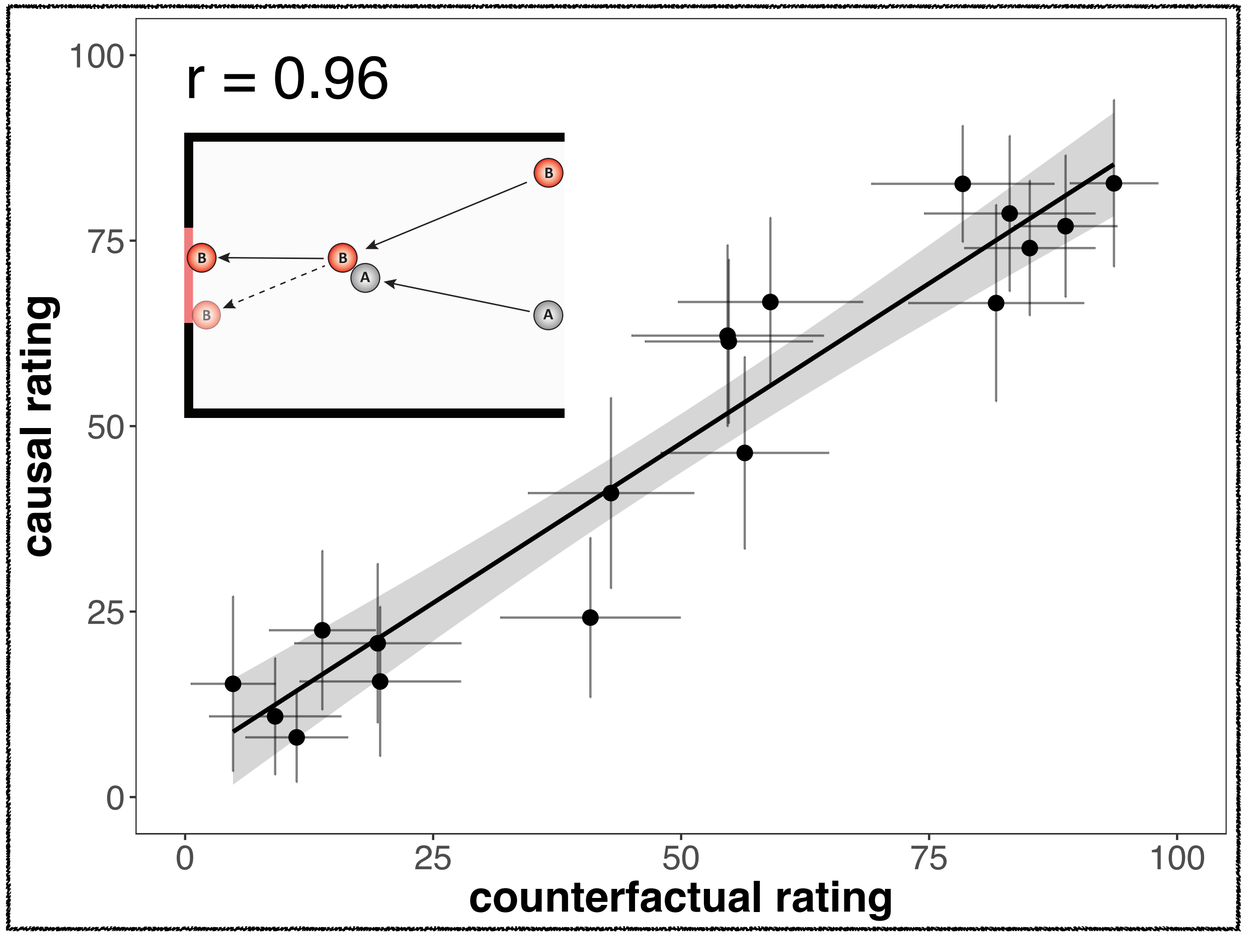


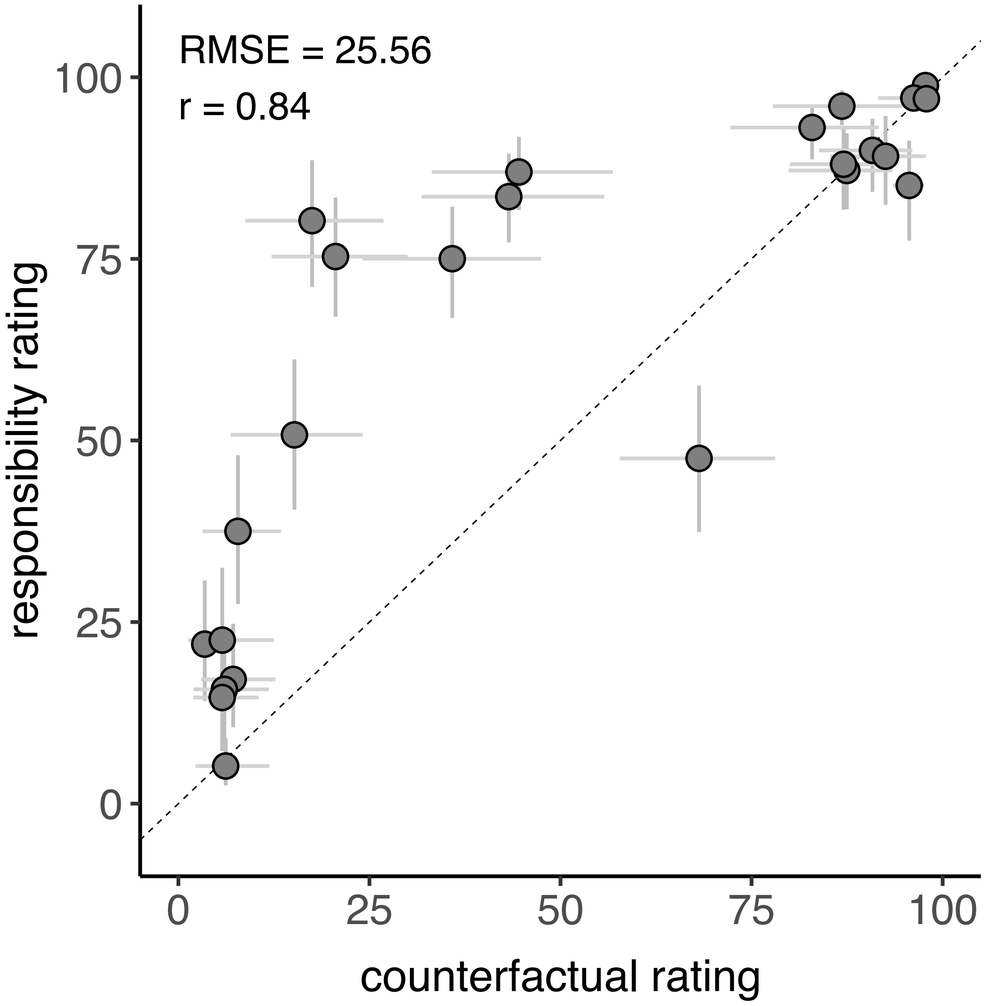

100
75
50
25
0
responsibility rating
0
25
50
75
100
counterfactual rating

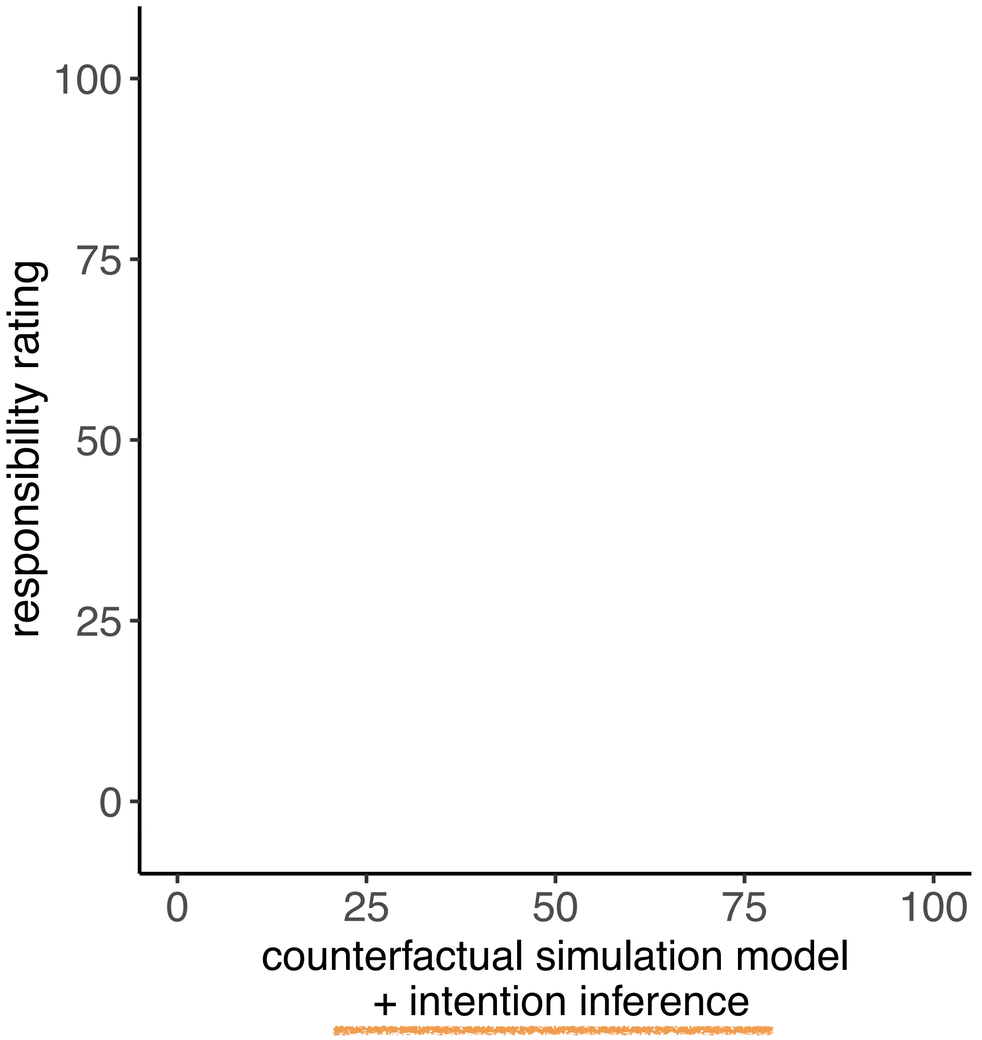

RMSE = 25.56
r = 0.84
responsibility rating
counterfactual rating
0
25
50
75
100

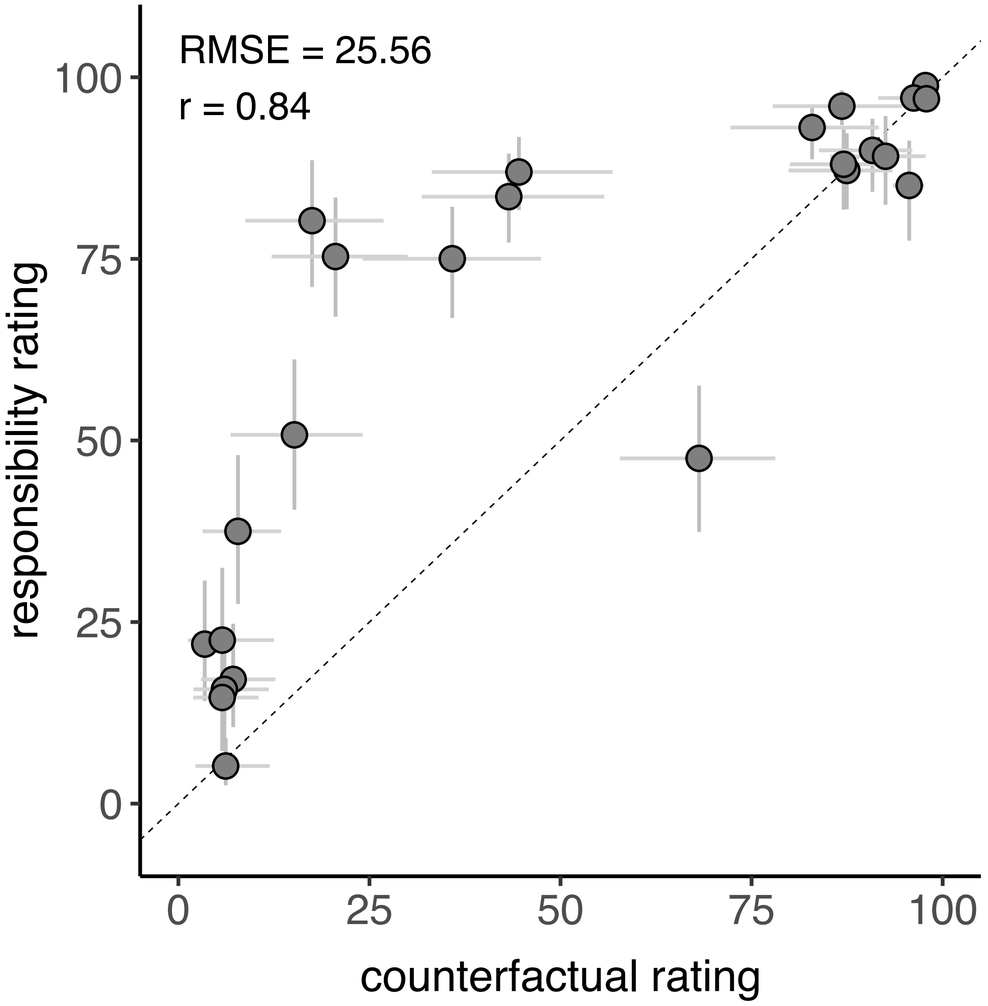


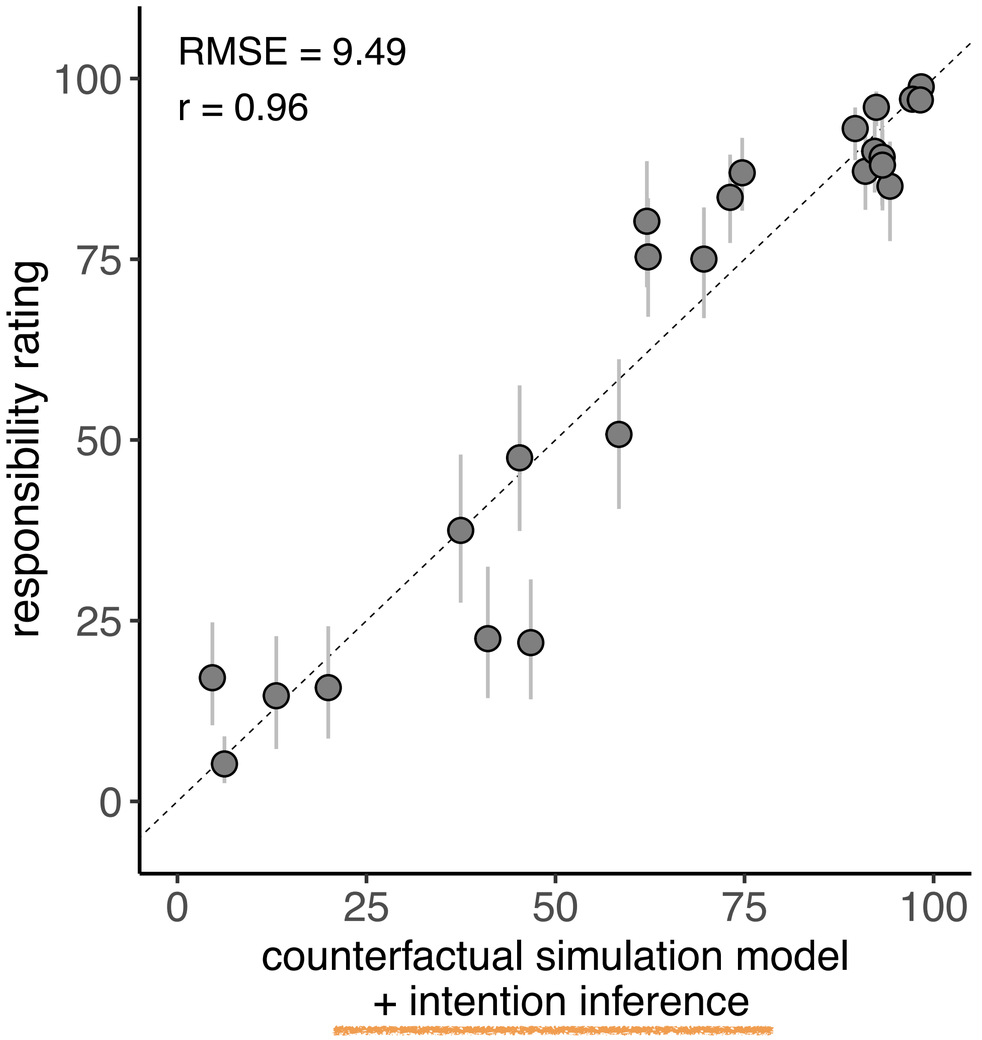
doesn't look like this

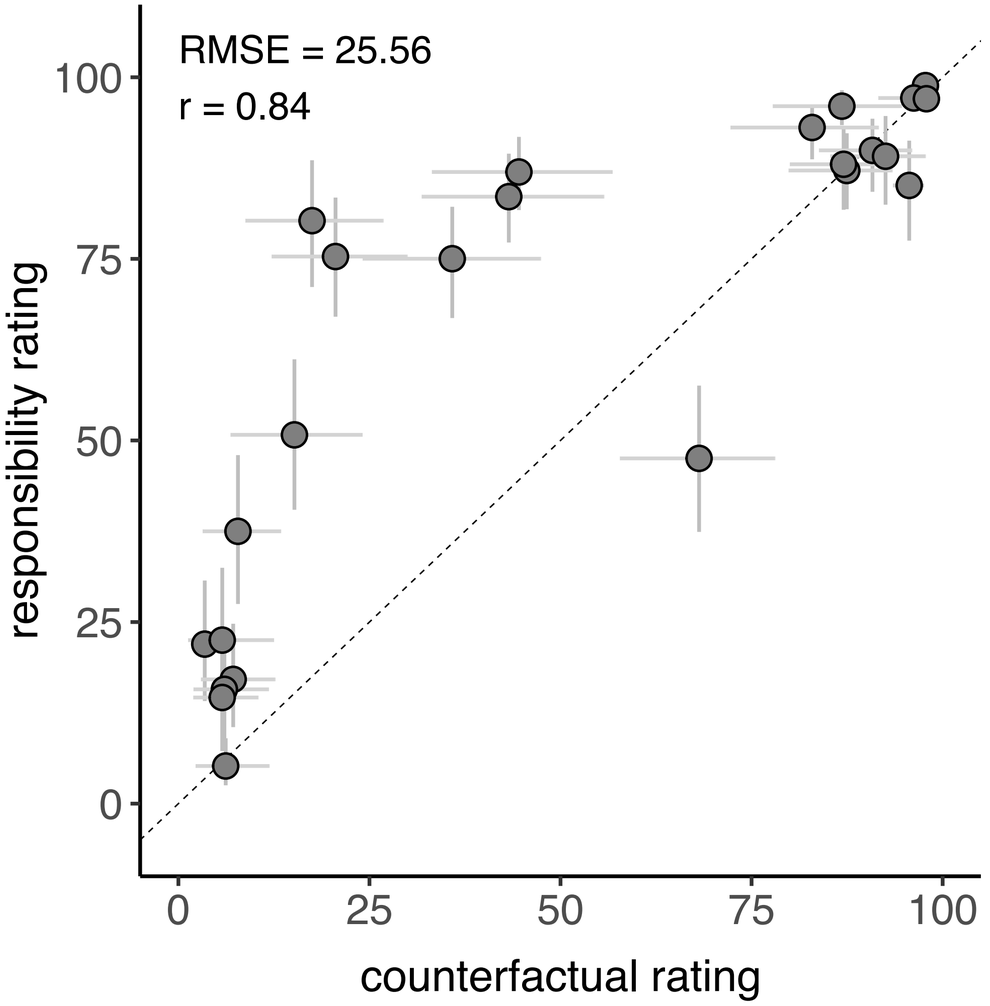


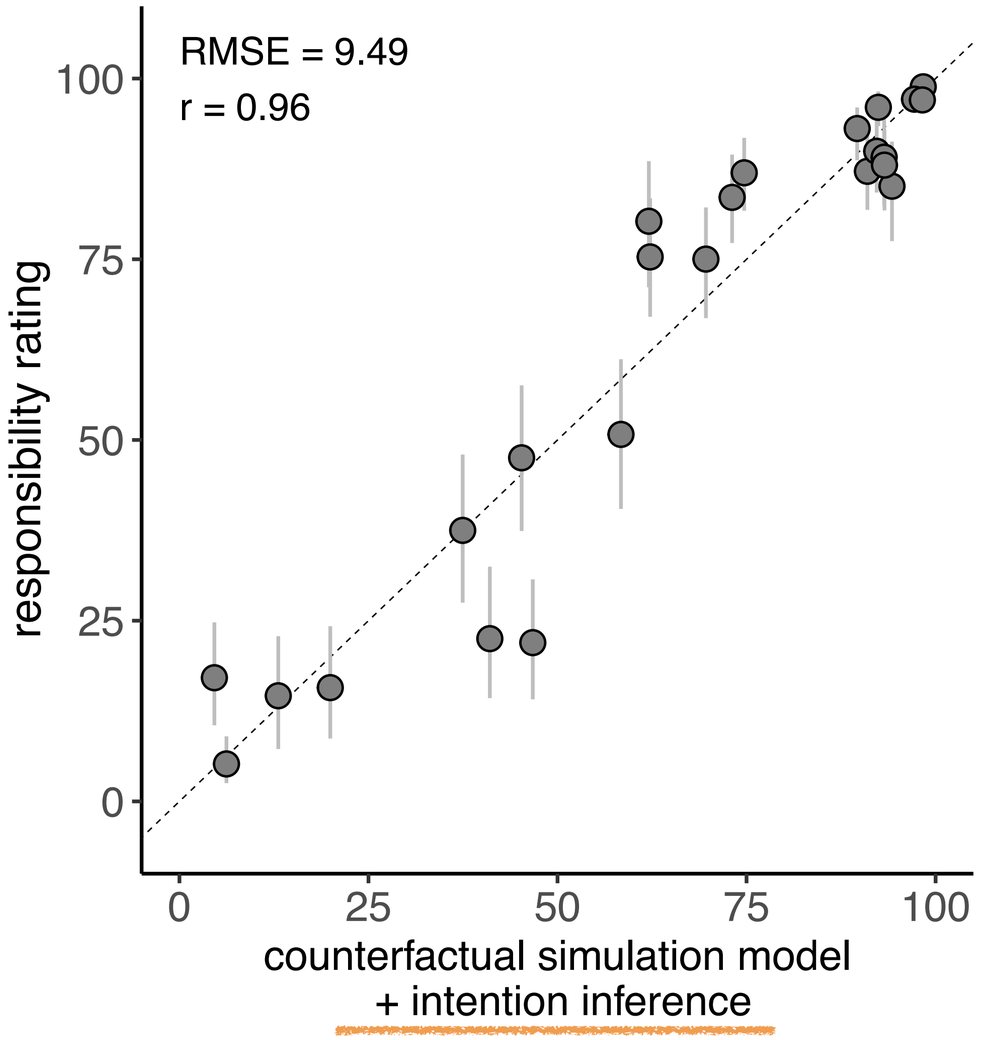

RMSE = 25.56
r = 0.84
responsibility rating
counterfactual rating
0
25
50
75
100


responsibility rating
counterfactual simulation model
+ intention inference
0
25
50
75
100


RMSE = 25.56
r = 0.84
responsibility rating
counterfactual rating
0
25
50
75
100


RMSE = 9.49
r = 0.96
responsibility rating
counterfactual simulation model
+ intention inference
0
25
50
75
100





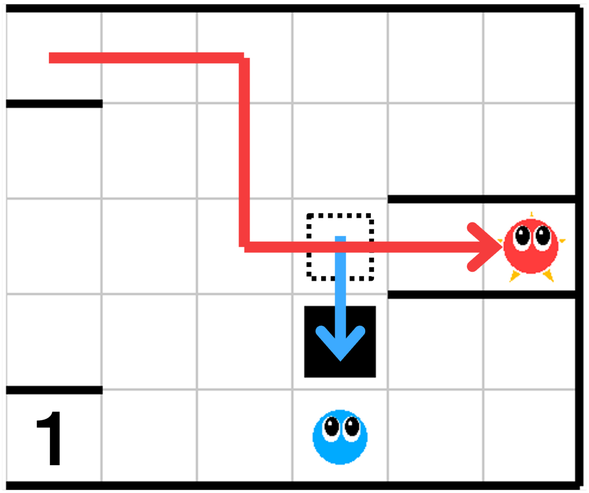
A model that combines

counterfactual simulation + intention inference

accurately captures responsibility judgments

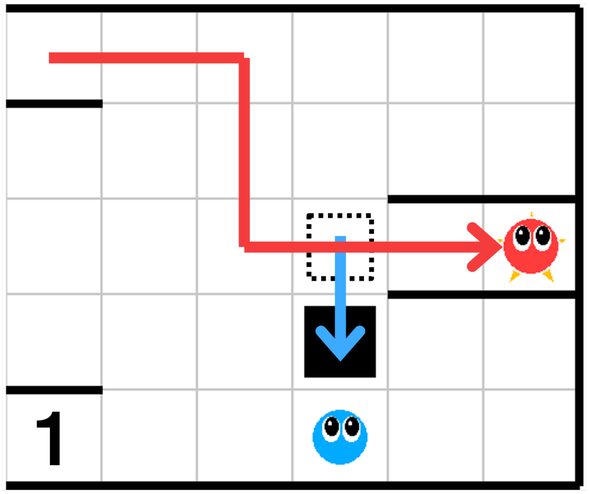
Counterfactual simulation & intuitive psychology

Counterfactual simulation & intuitive psychology

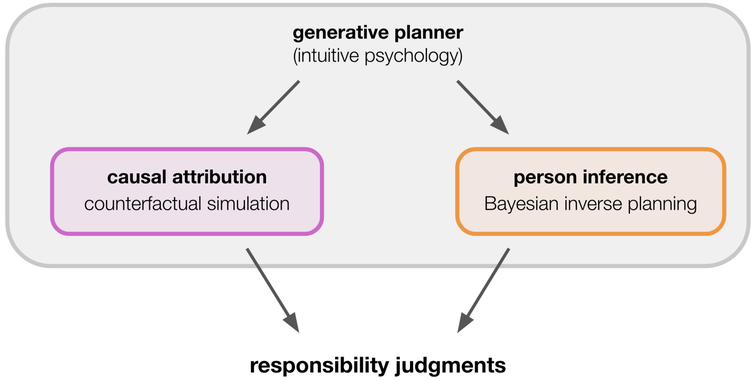


Judging whether someone **helped or hindered** requires counterfactual simulation

Counterfactual simulation & intuitive psychology



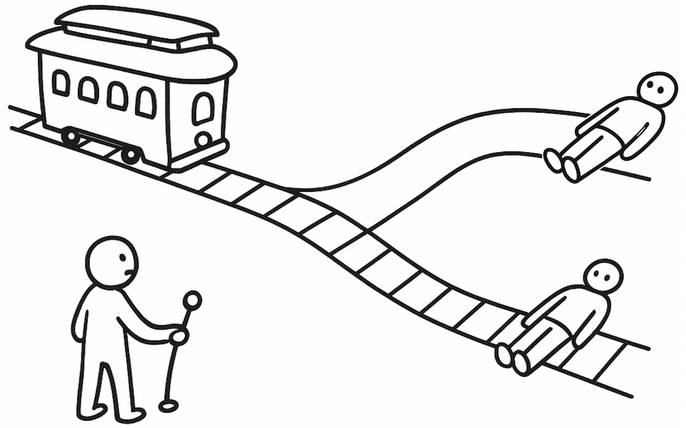
Judging whether someone **helped or hindered** requires counterfactual simulation



Responsibility judgments are sensitive to the agent's **causal role** and their **inferred mental states**

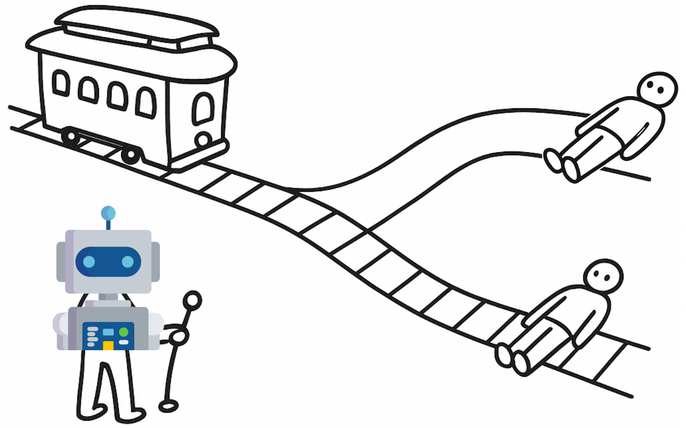
Are counterfactuals relevant for AI?

Are counterfactuals relevant for AI?

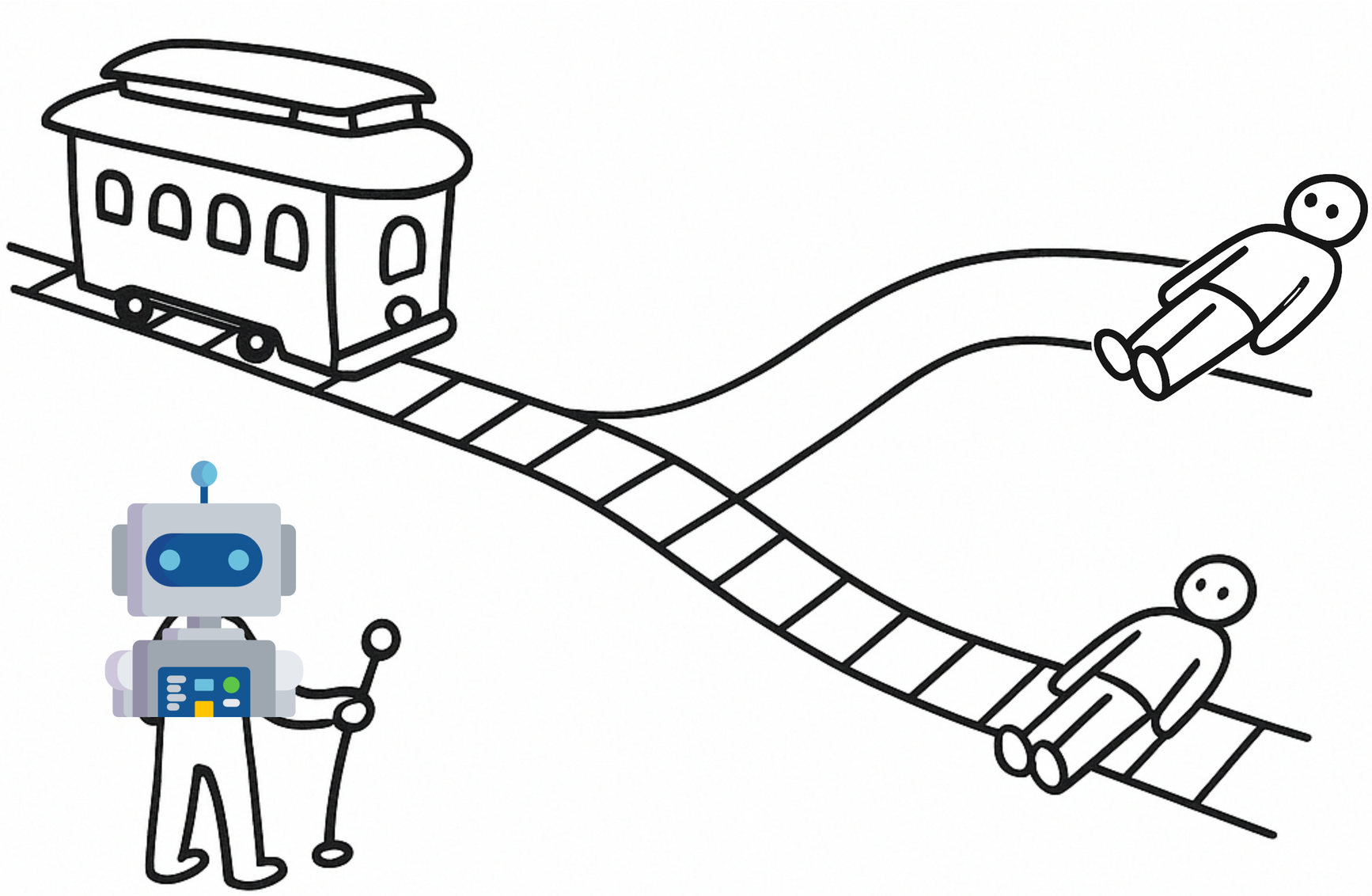


trolley dilemma

Are counterfactuals relevant for AI?



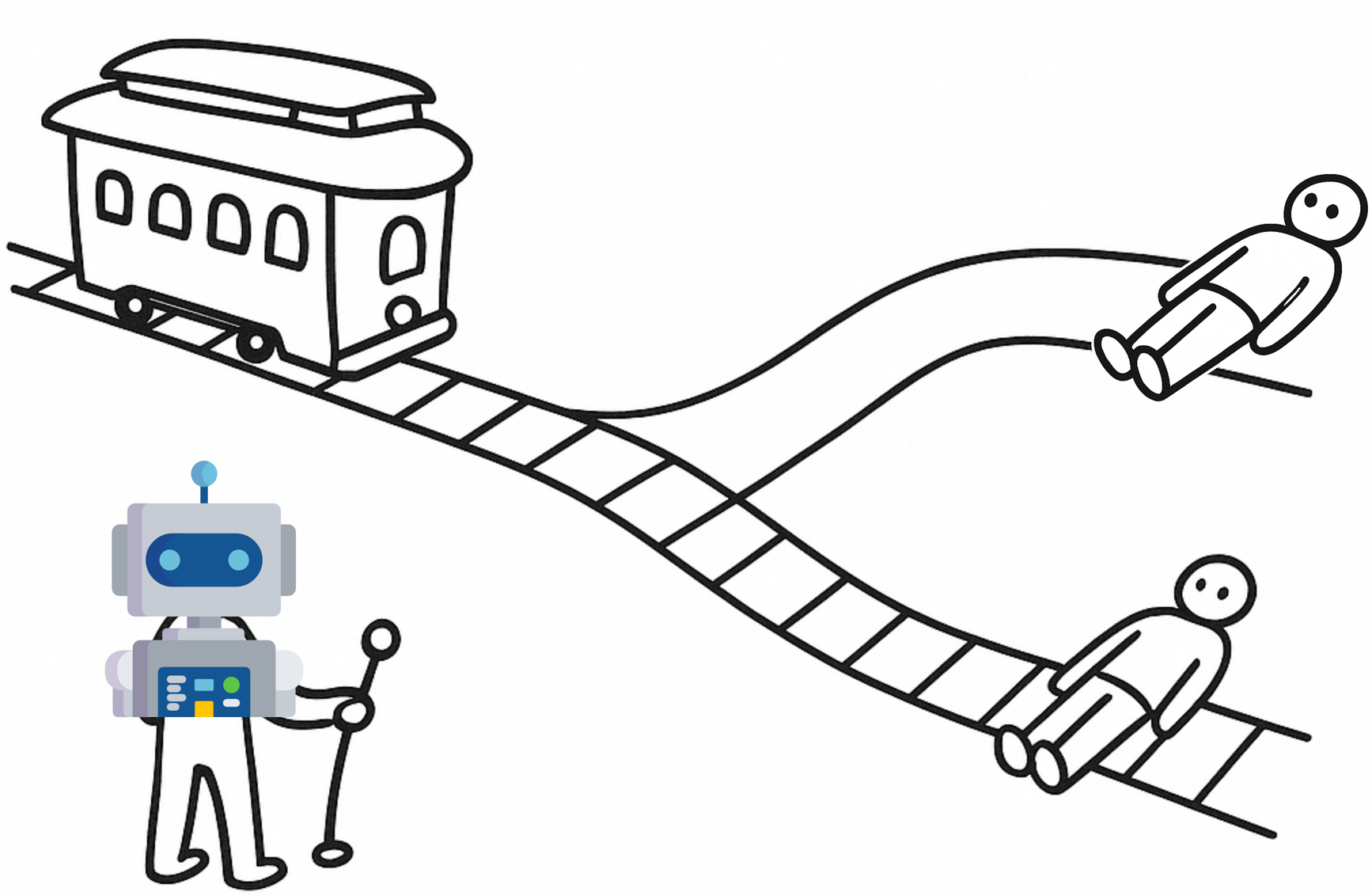
trolley dilemma



1

0

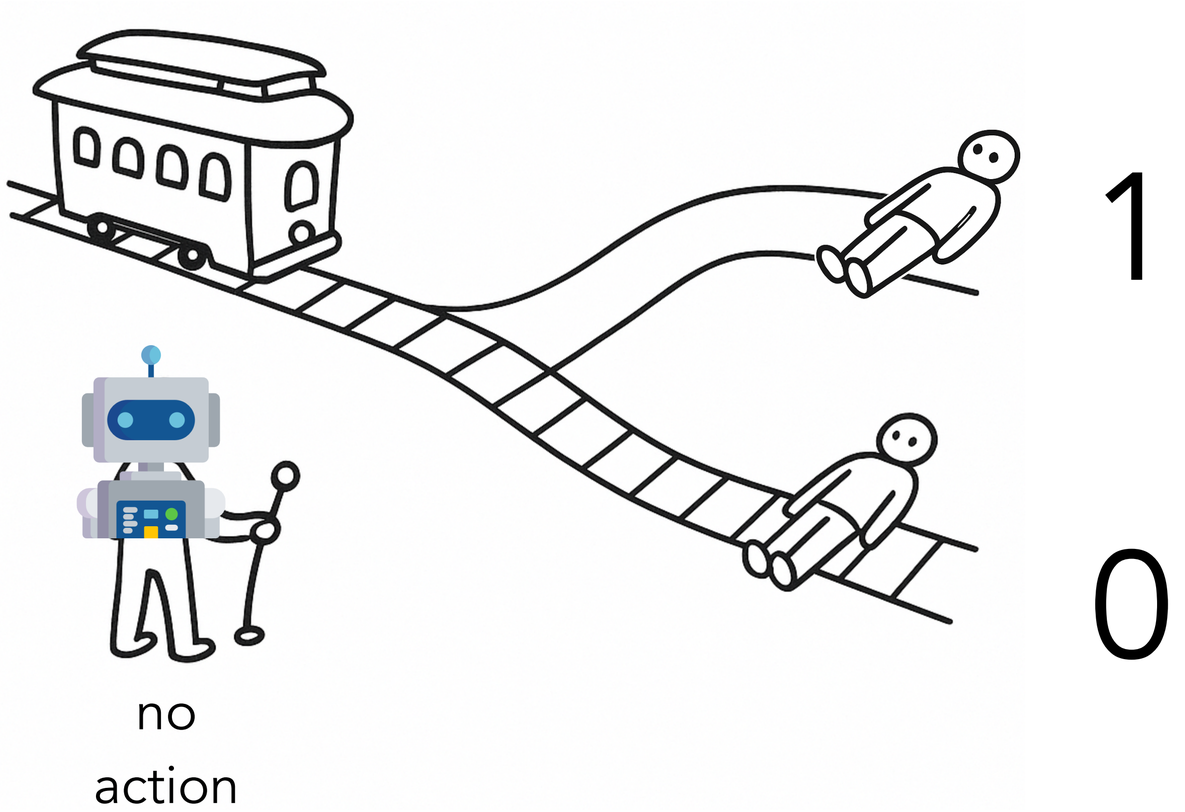
no action



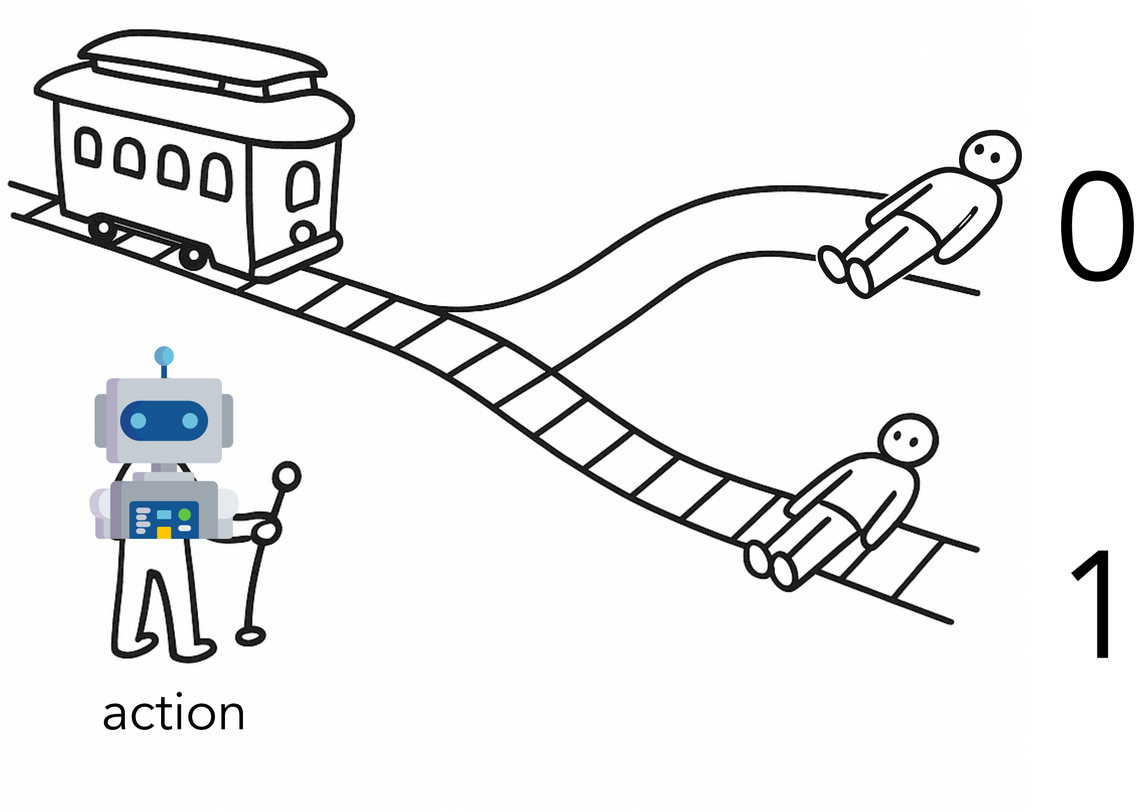
0

1

action

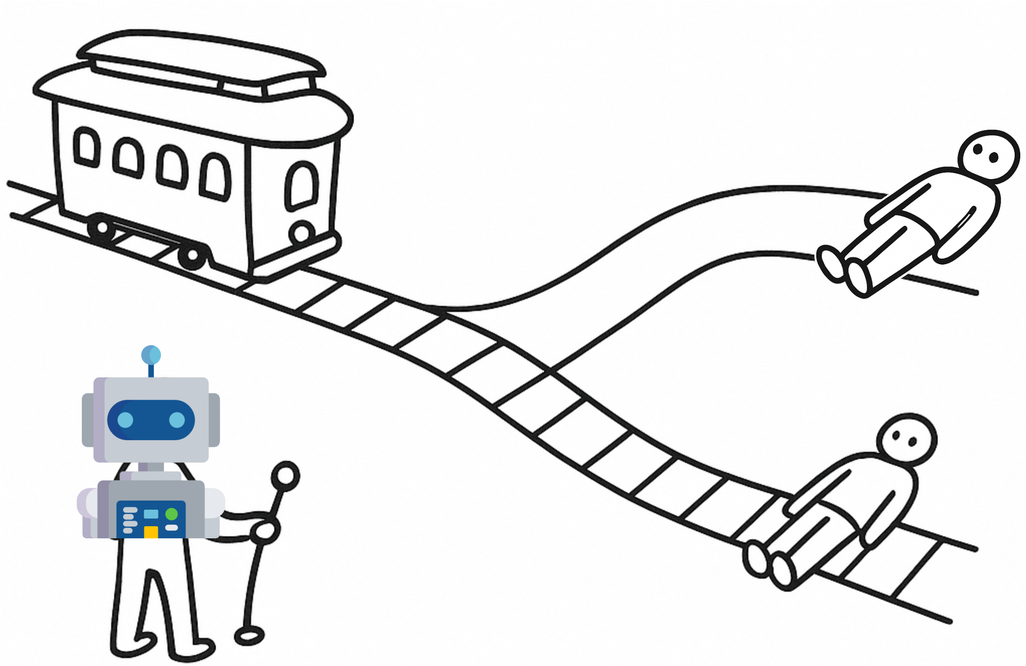


no
action

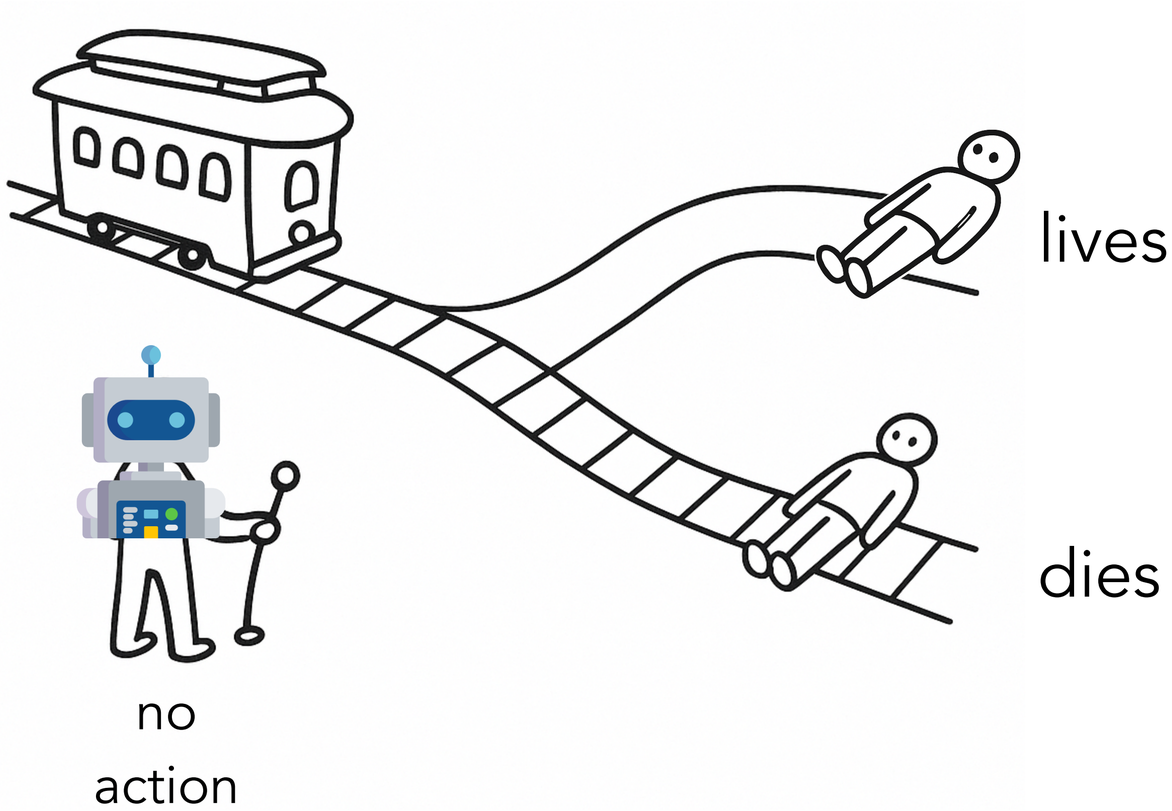


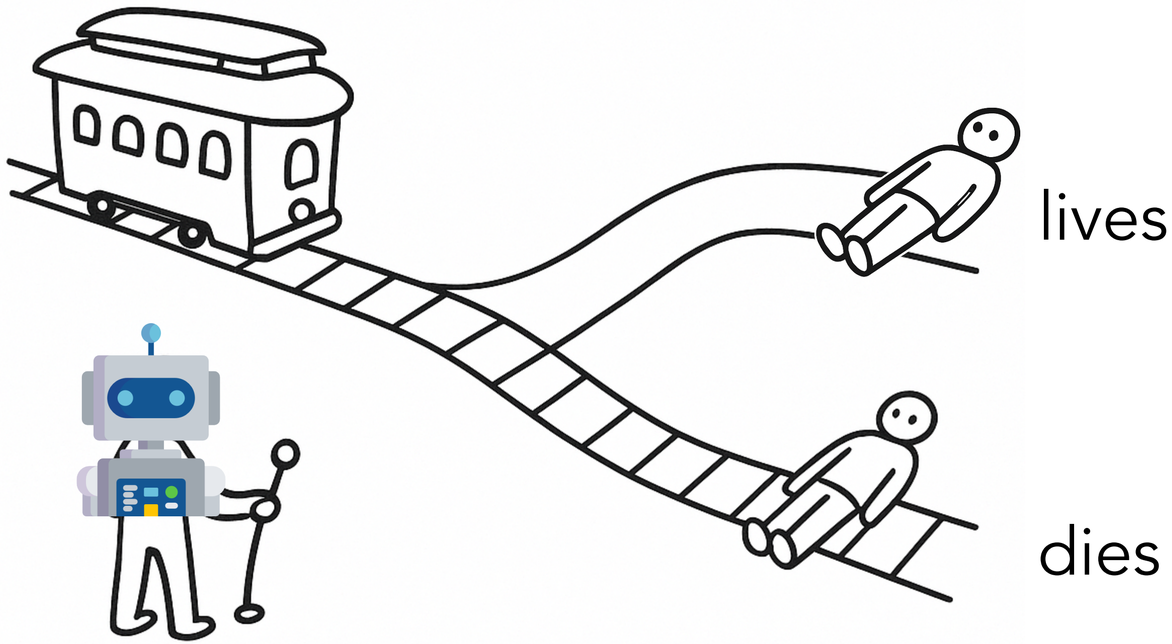
action

It makes no difference whether the AI acts

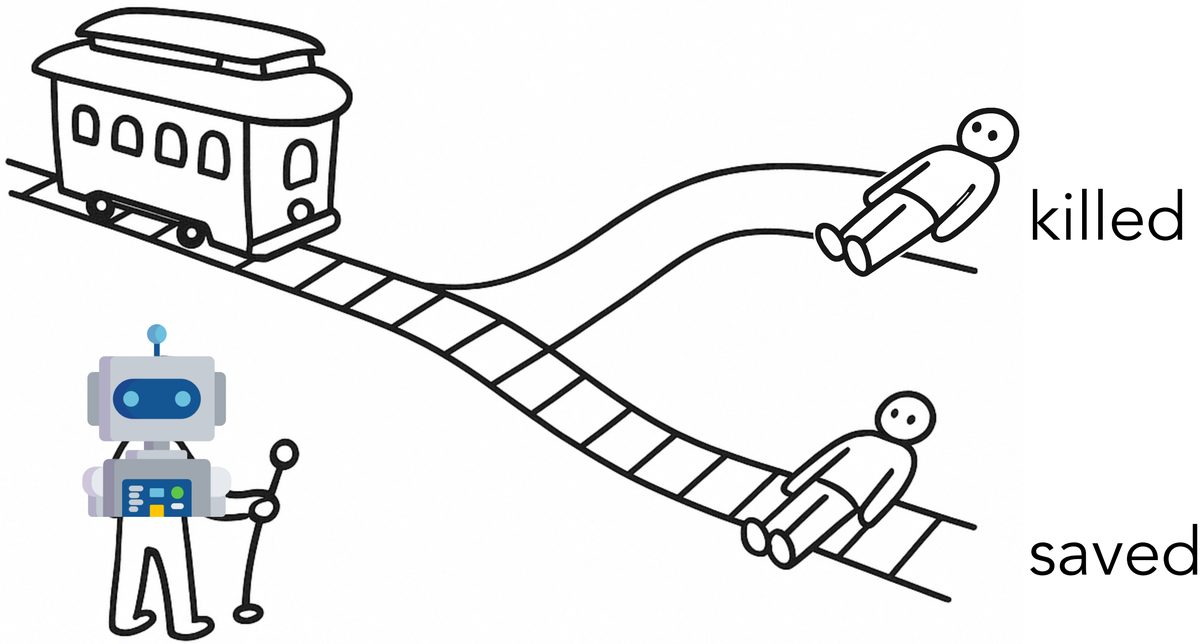


no
action

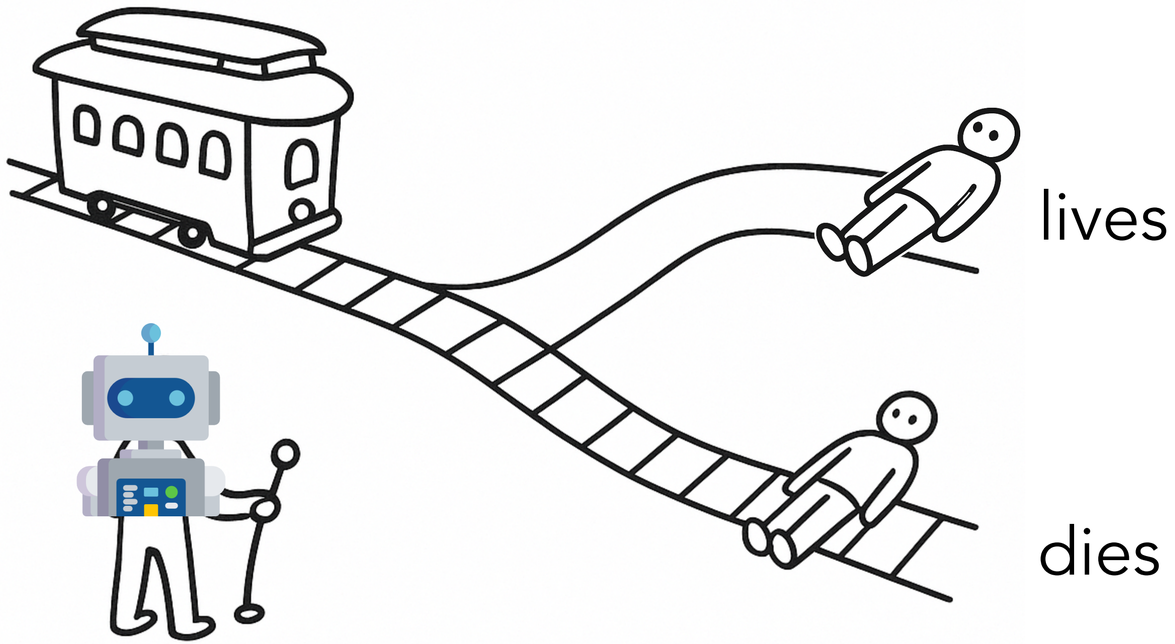

lives
dies
no
action

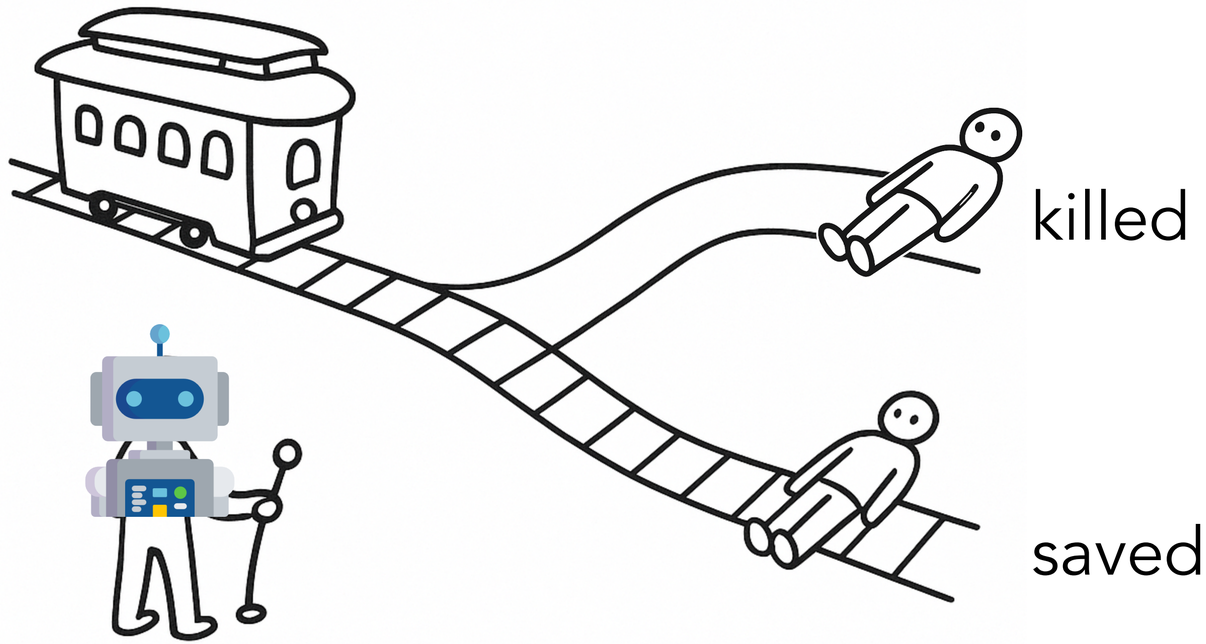


no action



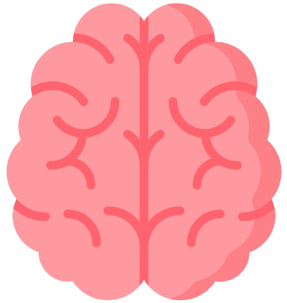
action





Saving someone is **good** but killing someone is **really bad**

Outline



Cognitive science

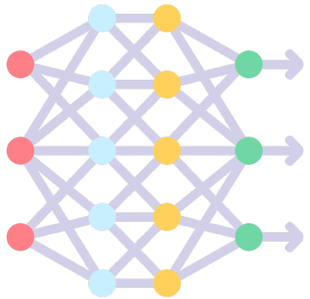
Machine learning



Large language models

Outline

Cognitive science



Machine learning



Large language models

Causal machine learning

The amount of work at the interface of causality and machine learning, often referred to as **causal machine learning**, has been increasing very rapidly.

Kaddour et al. *"Causal machine learning: A survey and open problems."* arXiv preprint, 2022.
Peters et al. *"Elements of causal inference: foundations and learning algorithms."* The MIT Press, 2017.

Causal machine learning

The amount of work at the interface of causality and machine learning, often referred to as **causal machine learning**, has been increasing very rapidly.

Causal machine learning operationalizes causal (counterfactual) reasoning about

the **outputs** of machine learning models,
the **data** used by these models, and
the **users** of these models

using the theoretical framework of **structural causal models (SCMs).**

Kaddour et al. *"Causal machine learning: A survey and open problems."* arXiv preprint, 2022.
Peters et al. *"Elements of causal inference: foundations and learning algorithms."* The MIT Press, 2017.

Structural Causal Models (SCMs)

Given a set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, a SCM defines a **complete data-generating process** via a collection of assignments

$$X_i := f_i(\mathbf{PA}_i, U_i),$$

where $\mathbf{PA}_i \subseteq \mathbf{X} \backslash X_i$ are the direct causes of X_i,

$\mathbf{U} = \{U_1, \ldots, U_n\}$ are jointly independent noise variables

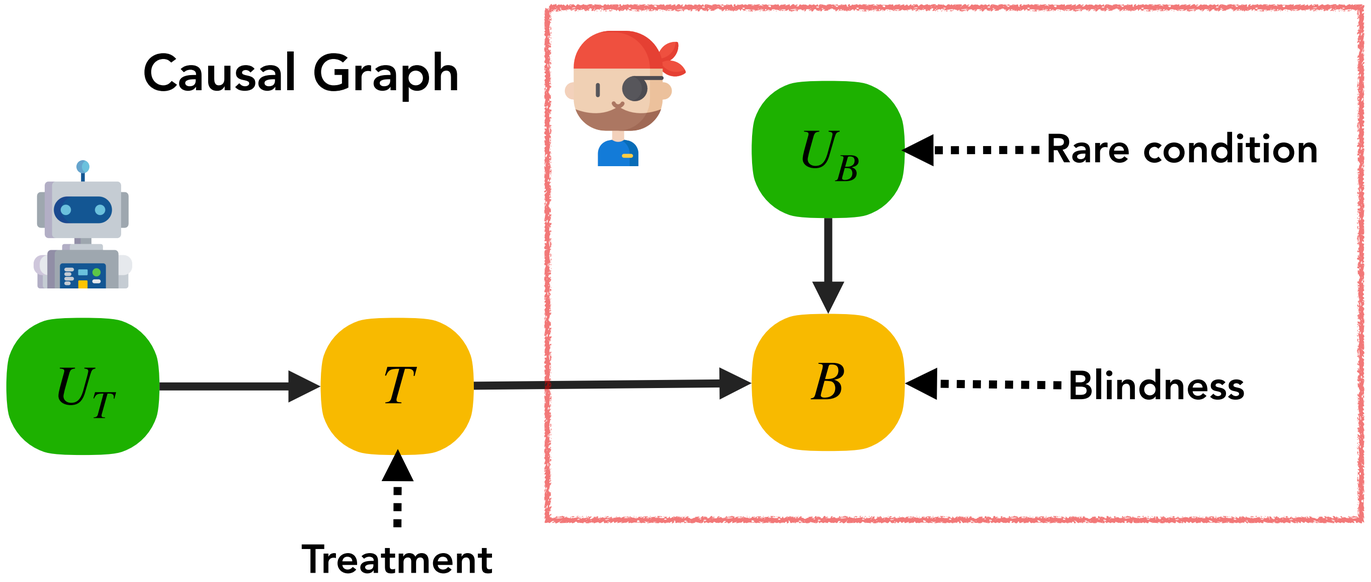
$\mathbf{F} = \{f_1, \ldots, f_n\}$ are deterministic causal mechanisms, and

$P(\mathbf{U})$ denotes the (prior) distribution of the noise variables.

Pearl. "*Causality.*" Cambridge university press, 2009.
Peters et al. *"Elements of causal inference: foundations and learning algorithms."* The MIT Press, 2017.

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) Interventional and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

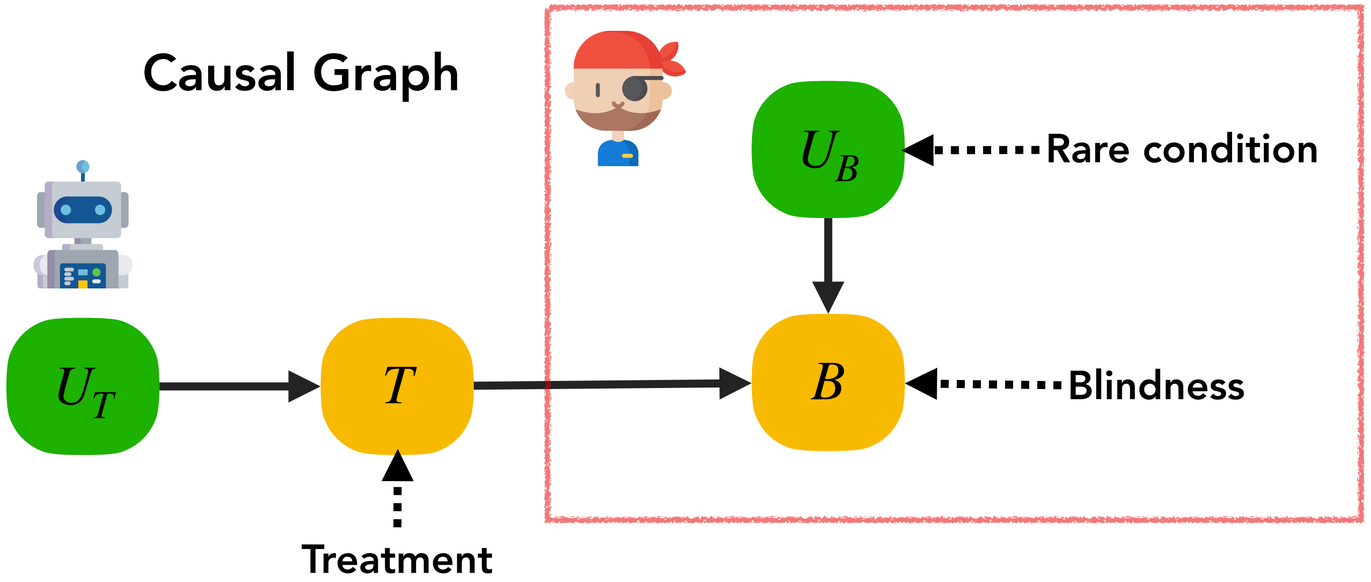
$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) **Observational**, (2) Interventional and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

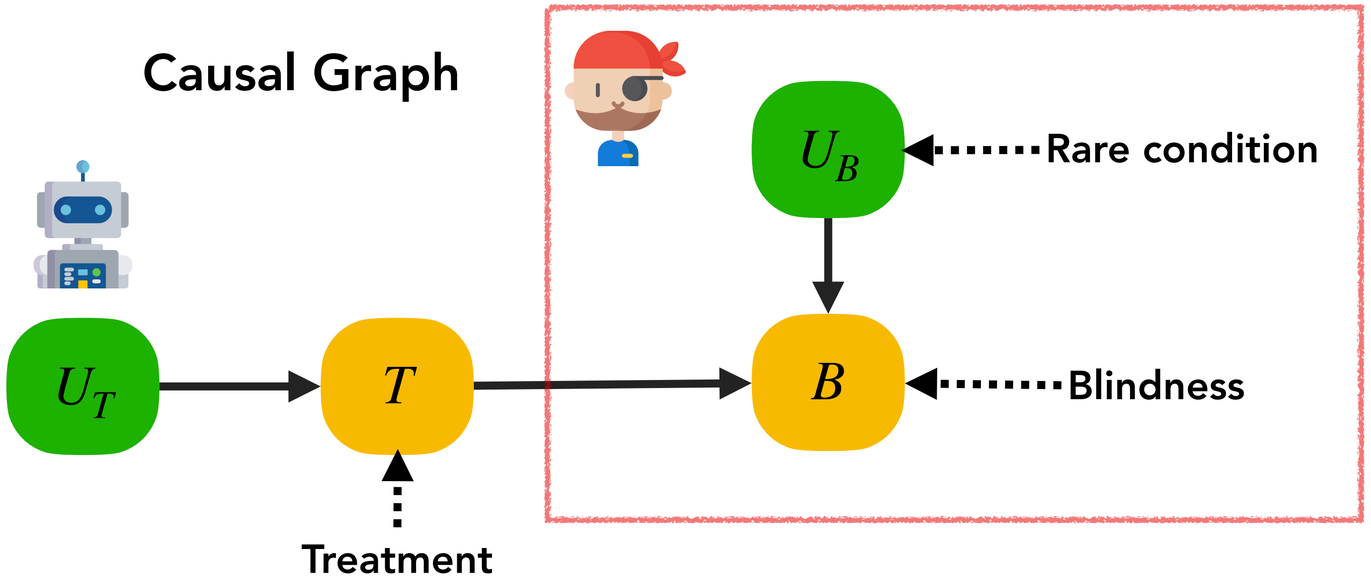
Observational question

What will happen to the patient?

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) **Observational**, (2) Interventional and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

"observe" →

Observational question

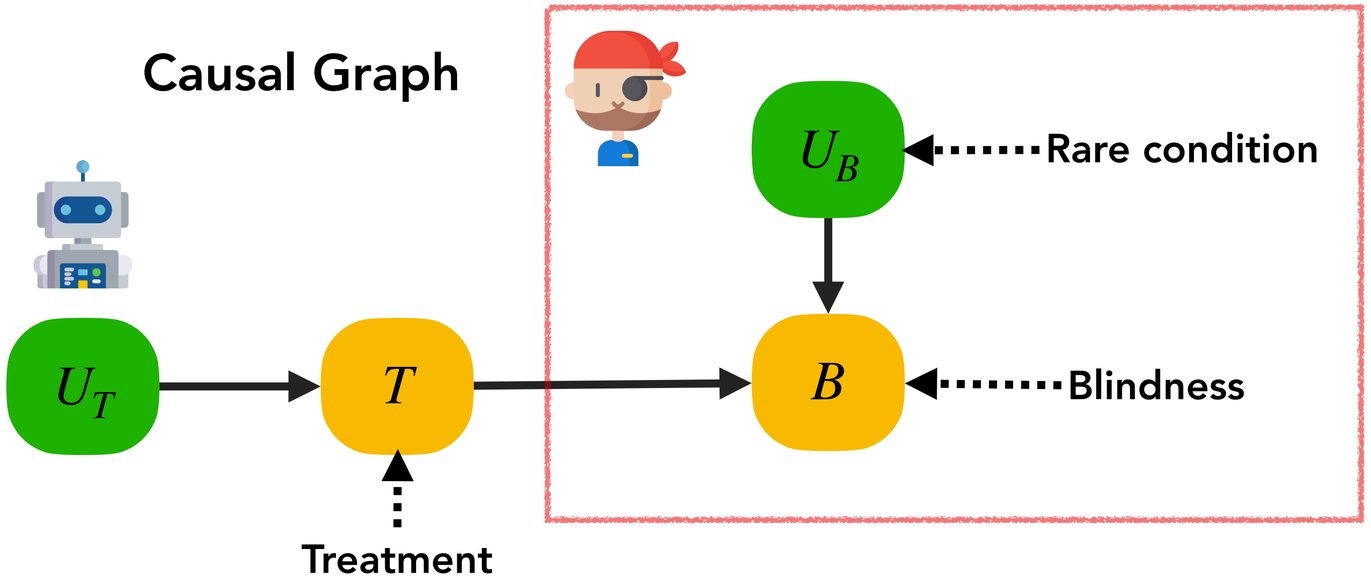
What will happen to the patient?

The patient will get blind ($B = 1$) with prob. 0.5

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) **Observational**, (2) Interventional and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

"observe" →

Observational question

What will happen to the patient?

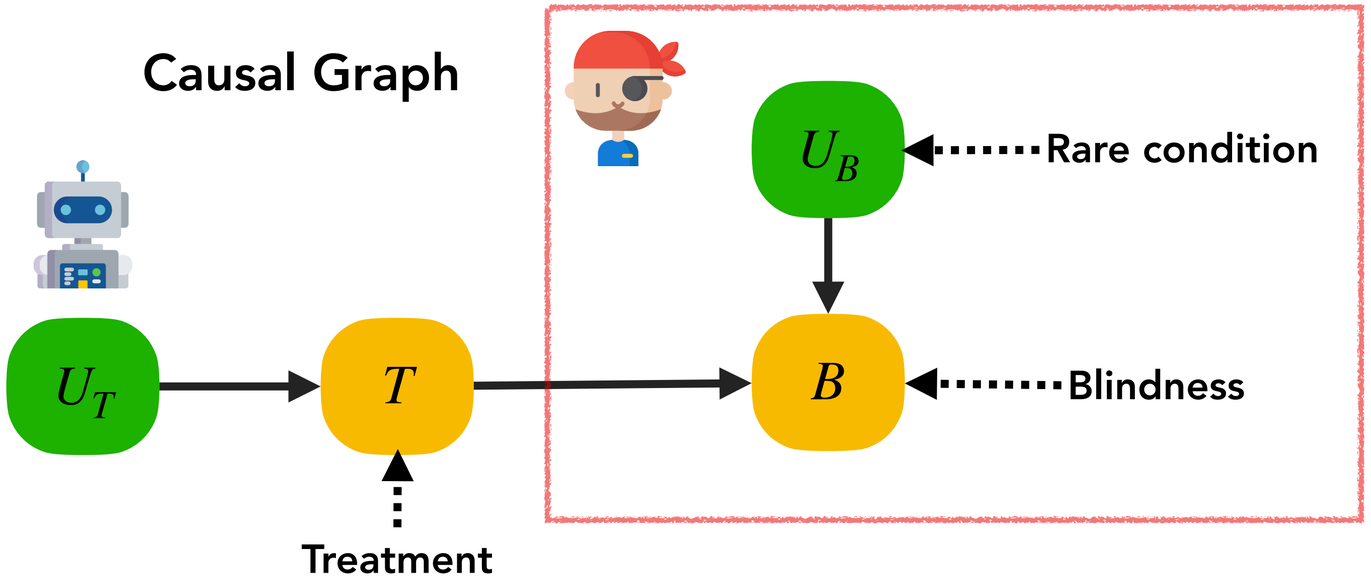
The patient will get blind ($B = 1$) with prob. 0.5

Formally, $P^{\mathscr{M}}(B = 1) = 0.5$

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) **Interventional** and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

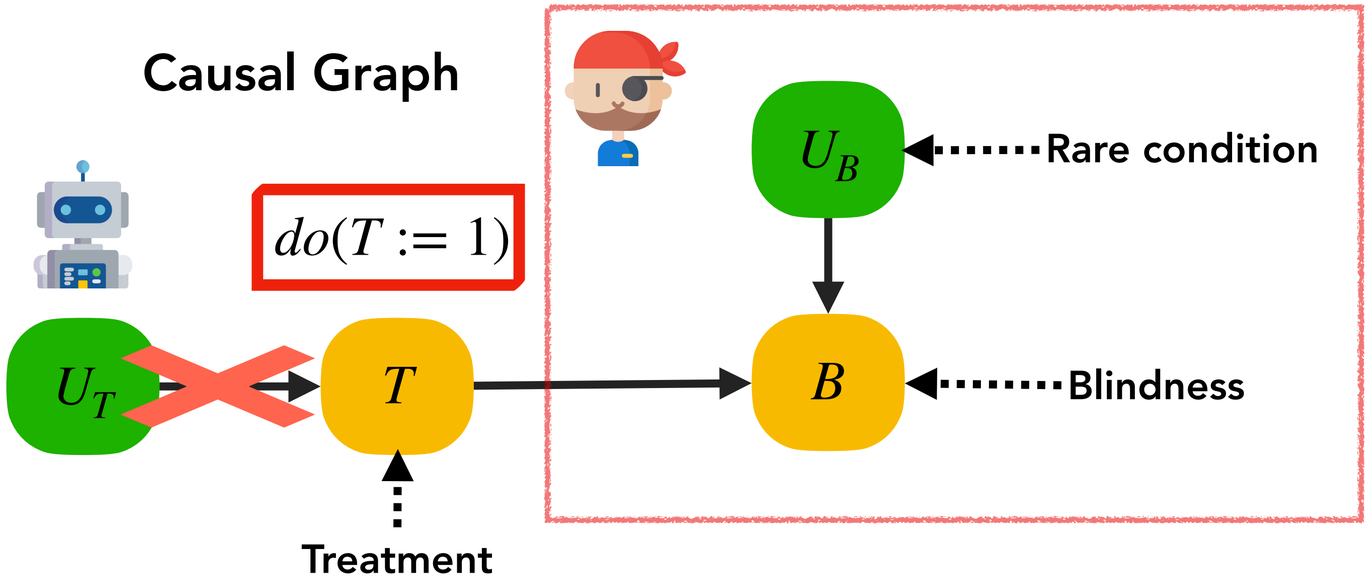
Interventional question

What will happen to the patient if a doctor breaks the robot and always administers the treatment?

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) **Interventional** and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

~~$T := f_T(U_T)$~~ $T := 1$

$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$

$U_B \sim Ber(0.01)$, ~~$U_T \sim Ber(0.5)$~~

"do"

Interventional question

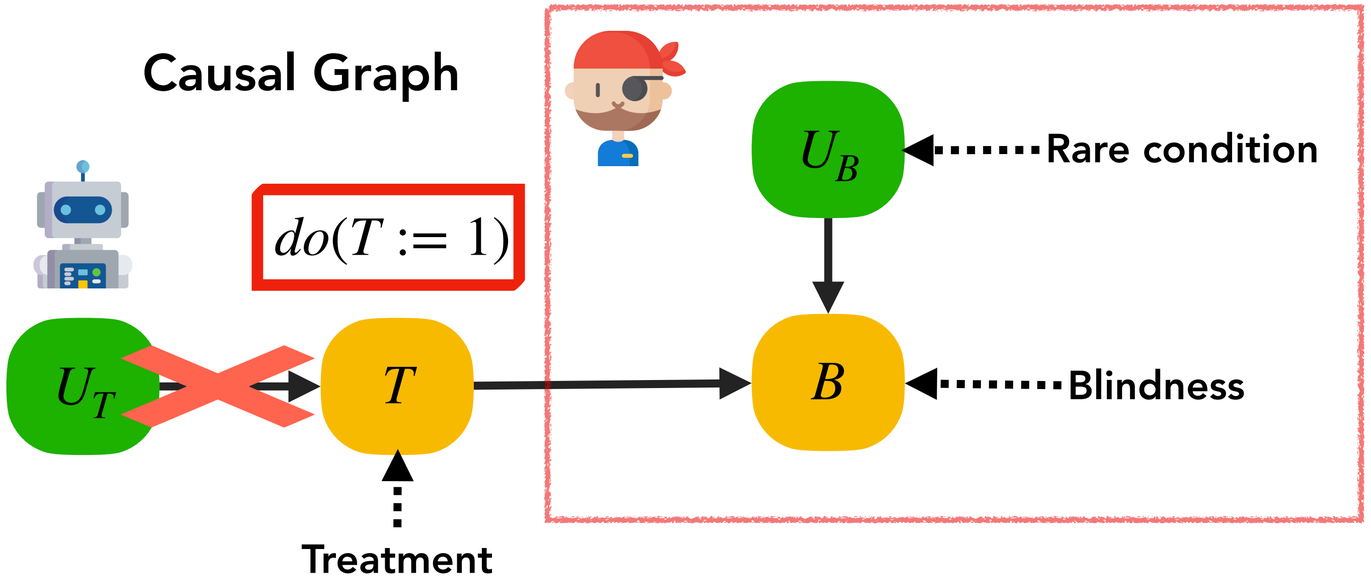
What will happen to the patient if a doctor breaks the robot and always administers the treatment?

The patient will get blind ($B = 1$) with prob. 0.01

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) **Interventional** and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$T := 1$

$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$

$U_B \sim Ber(0.01)$,

"do"

Interventional question

What will happen to the patient if a doctor breaks the robot and always administers the treatment?

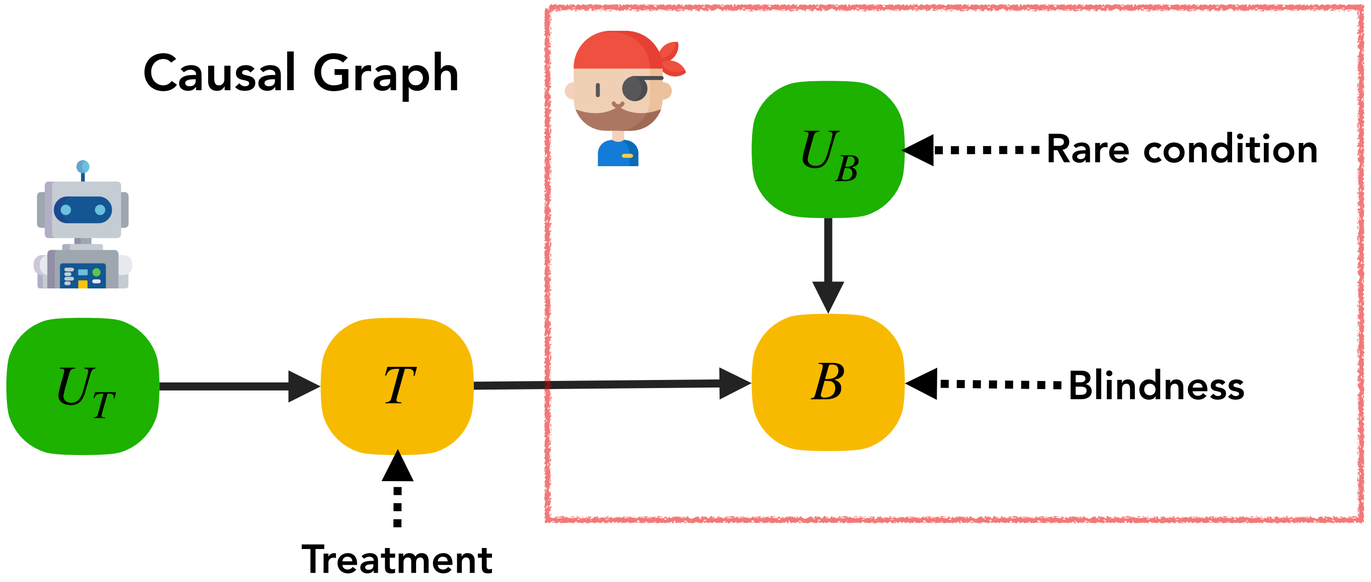
The patient will get blind ($B = 1$) with prob. 0.01

Formally, $P^{\mathscr{M};\, do(T=1)}(B = 1) = 0.01$

Example adapted from *Elements of causal inference*, MIT Press, 2017

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) Interventional and (3) **Counterfactual** Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

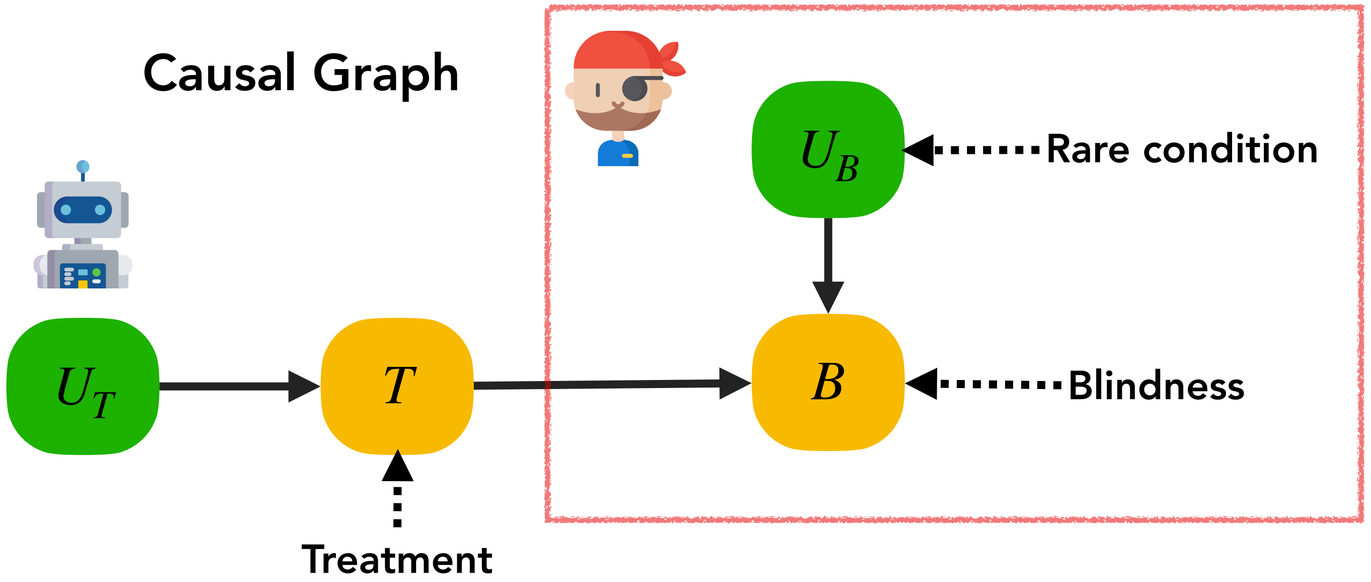
Counterfactual question

The treatment was administered and the patient got blind. What would have happened if the treatment had not been administered?

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) Interventional and (3) **Counterfactual** Queries



Modified Structural Causal Model $\mathscr{M}_{T=1,B=1}$

$T := 1$

$B := T$

$U_B = 1$ with prob. 1 ◄- - - **Posterior distribution of the noise**

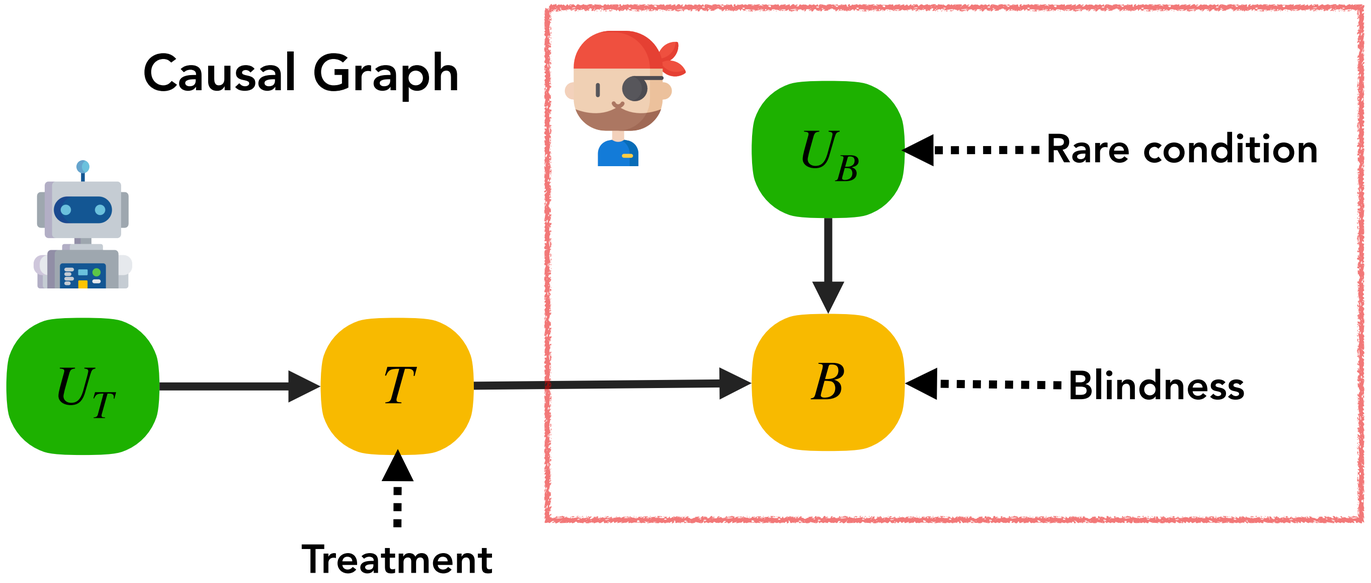
Counterfactual question

The treatment was administered and the patient got blind. What would have happened if the treatment had not been administered?

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) Interventional and (3) **Counterfactual** Queries



Modified Structural Causal Model $\mathscr{M}_{T=1,B=1}$

~~$T := 1$~~ $T := 0$

$B := T$

$U_B = 1$ with prob. 1

"imagine"

Example adapted from *Elements of causal inference, MIT Press, 2017*

Counterfactual question

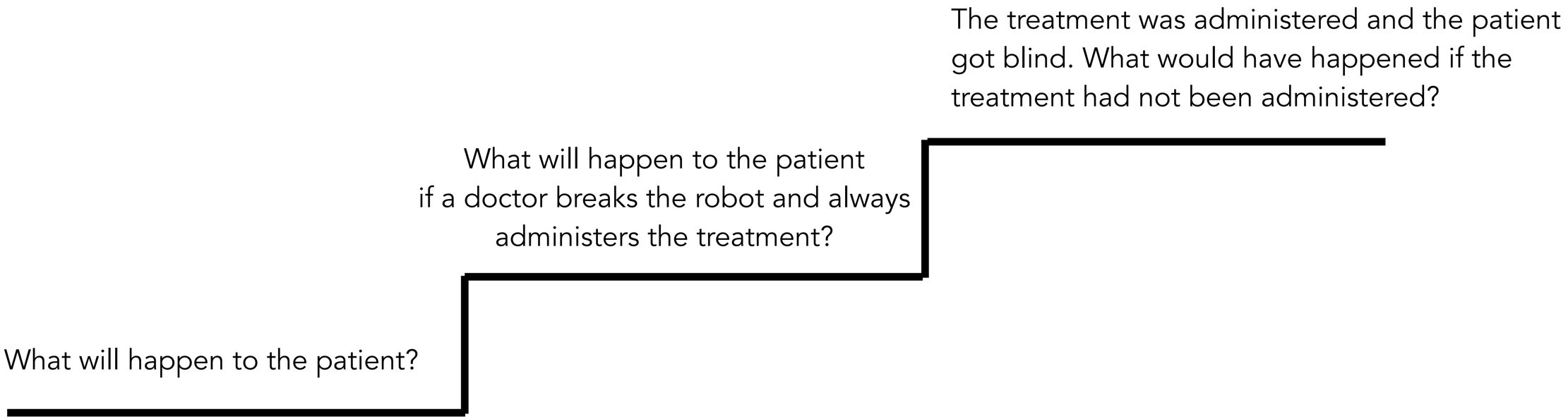
The treatment was administered and the patient got blind. What would have happened if the treatment had not been administered?

The patient would not have gotten blind ($B = 0$)

Formally, $P^{\mathscr{M}\,|\,T=1,B=1\,;\,do(T=1)}(B = 1) = 0$

The ladder of causation

(1) Observational, (2) Interventional and (3) Counterfactual Queries



It is called **ladder of causation** because questions at level $i \in \{1,2,3\}$ can only be answered if information from level $j \geq i$ is available. Counterfactuals sit at the top of the ladder!

Pearl. "*Causality.*" Cambridge university press, 2009.
Bareinboim et al. "*On Pearl's hierarchy and the foundations of causal inference.*" Probabilistic and causal inference: the works of Judea Pearl, 2022.

Identifiability

Identification of

an interventional probability, e.g., $P^{\mathscr{M}\,;\,do(T=1)}(B)$, or

a counterfactual probability, e.g., $P^{\mathscr{M}\,|\,T=1,B=1\,;\,do(T=1)}(B)$

refers to the process of estimating it using (observational) data from $\mathscr{M}$.

Shpitser and Pearl. "*Complete identification methods for the causal hierarchy.*" JMLR, 2008.
Perkovic et al. "*Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs.*" JMLR, 2018.
Shalit et al. "*Estimating individual treatment effect: generalization bounds and algorithms.*" ICML, 2017.
Kallus. "*Treatment effect risk: Bounds and inference.*" Management Science, 2023.

Identifiability

Identification of

an interventional probability, e.g., $P^{\mathscr{M}\,;\,do(T=1)}(B)$, or

a counterfactual probability, e.g., $P^{\mathscr{M}\,|\,T=1,B=1\,;\,do(T=1)}(B)$

refers to the process of estimating it using (observational) data from $\mathscr{M}$.

If an interventional or counterfactual probability is not identifiable, then regardless of how much data we have, we will not be able to estimate it.

Shpitser and Pearl. "*Complete identification methods for the causal hierarchy.*" JMLR, 2008.
Perkovic et al. "*Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs.*" JMLR, 2018.
Shalit et al. "*Estimating individual treatment effect: generalization bounds and algorithms.*" ICML, 2017.
Kallus. "*Treatment effect risk: Bounds and inference.*" Management Science, 2023.

Identifiability

Identification of

an interventional probability, e.g., $P^{\mathscr{M}\,;\,do(T=1)}(B)$, or

a counterfactual probability, e.g., $P^{\mathscr{M}\,|\,T=1,B=1\,;\,do(T=1)}(B)$

refers to the process of estimating it using (observational) data from $\mathscr{M}$.

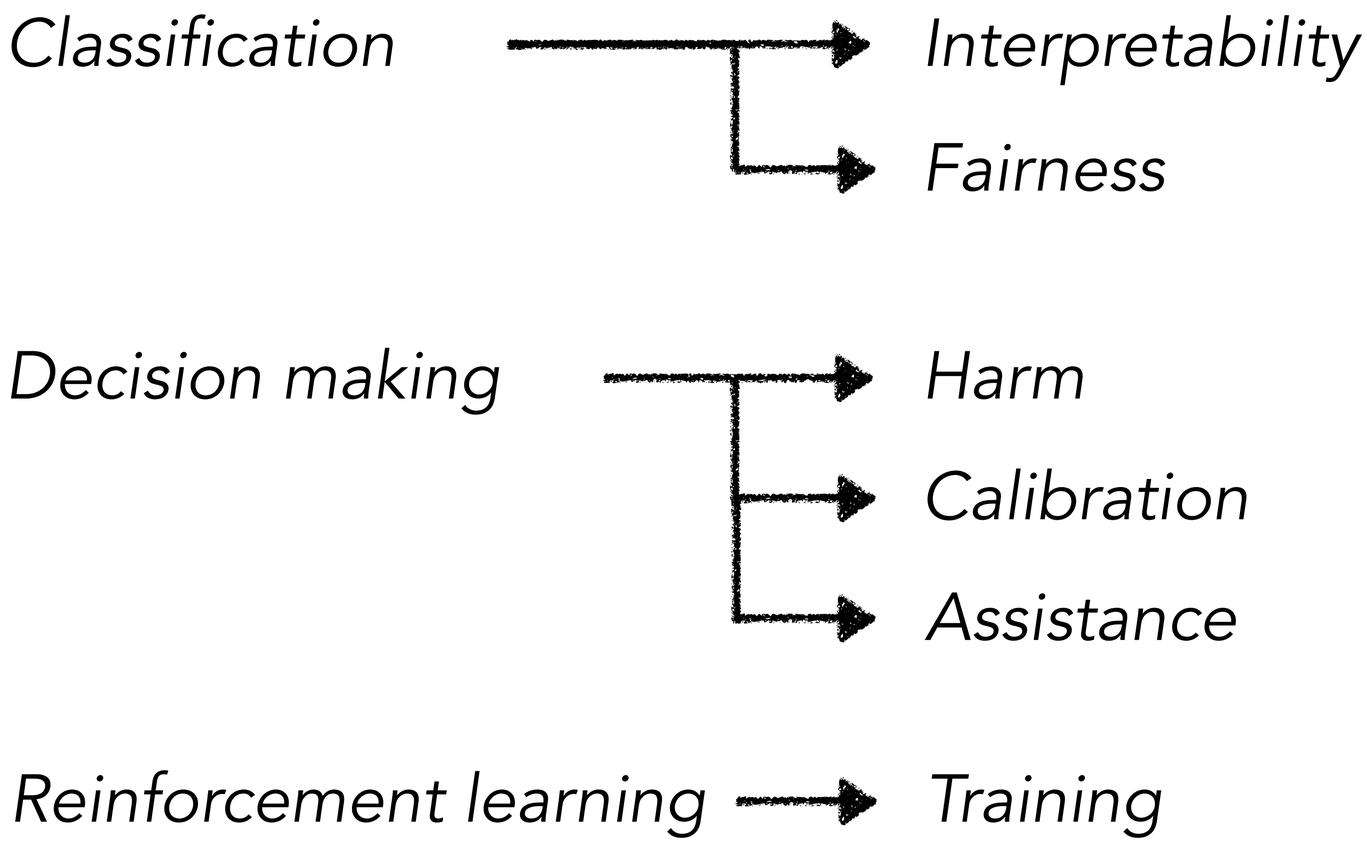
If an interventional or counterfactual probability is not identifiable, then regardless of how much data we have, we will not be able to estimate it.

There exist methods to

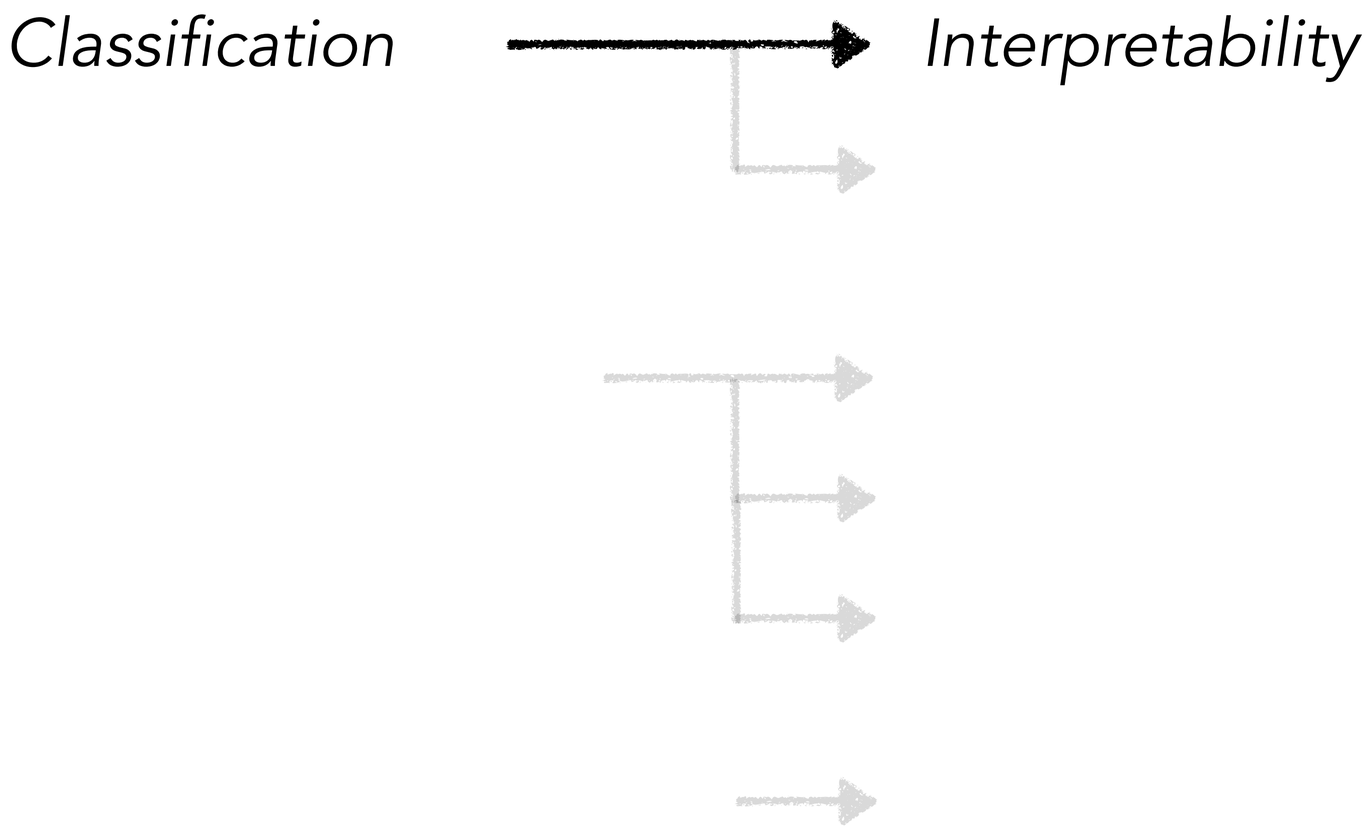
(i) determine the identifiability of specific interventional and counterfactual probabilities, and
(ii) estimate (or bound) quantities derived from these probabilities (e.g., individual treatment effects)

Shpitser and Pearl. "*Complete identification methods for the causal hierarchy.*" JMLR, 2008.
Perkovic et al. "*Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs.*" JMLR, 2018.
Shalit et al. "*Estimating individual treatment effect: generalization bounds and algorithms.*" ICML, 2017.
Kallus. "*Treatment effect risk: Bounds and inference.*" Management Science, 2023.

Use cases of counterfactuals in machine learning



Use cases of counterfactuals in machine learning



Counterfactual explanations

The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.

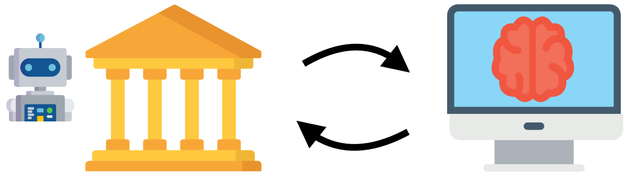
Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.



Kate



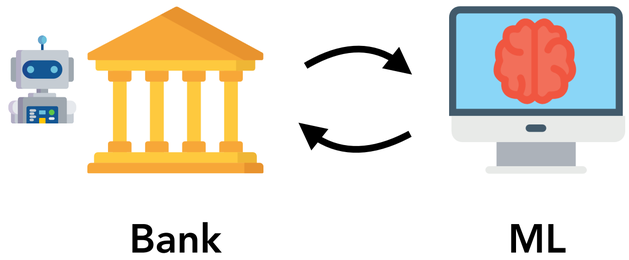
Bank ML

Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.

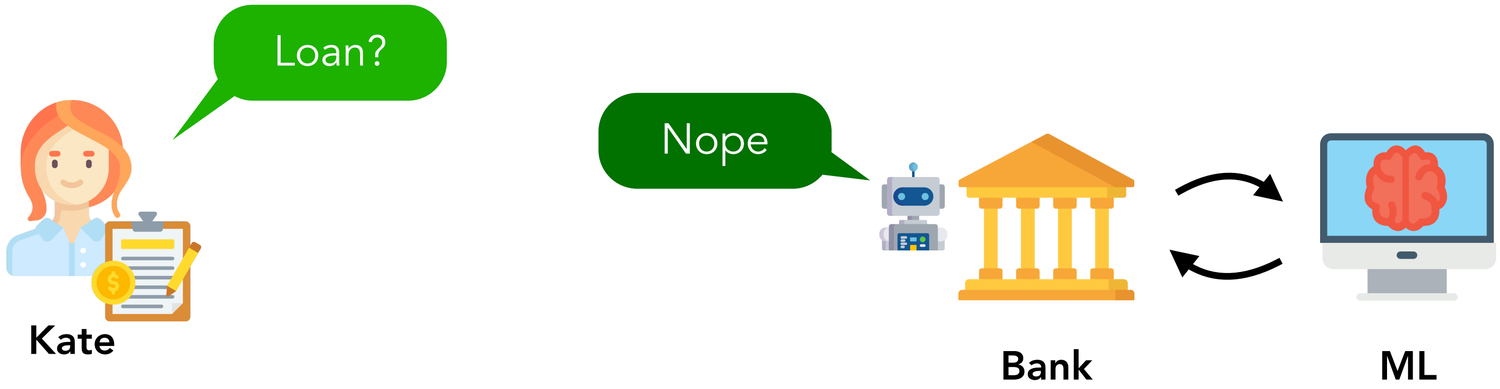




Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

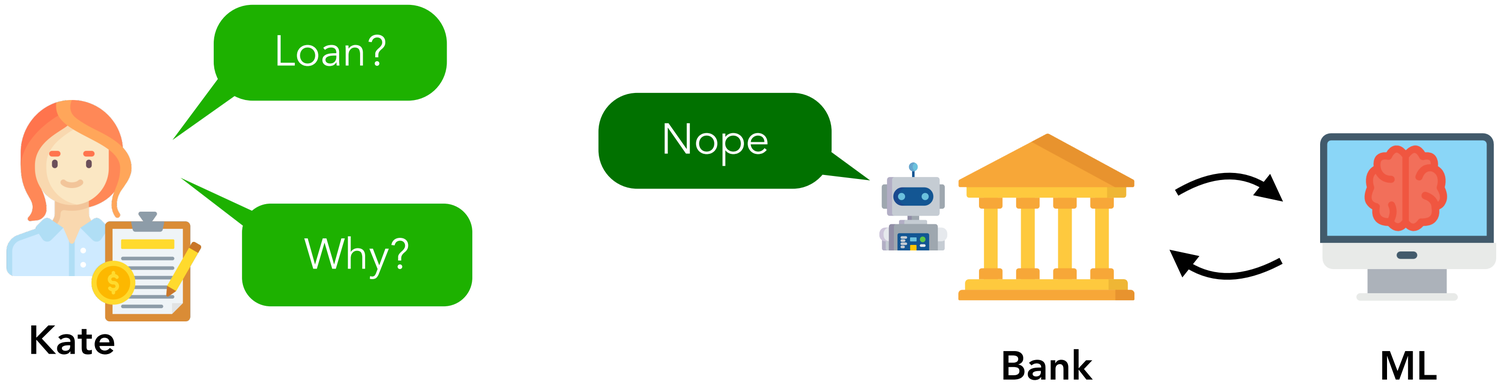
The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

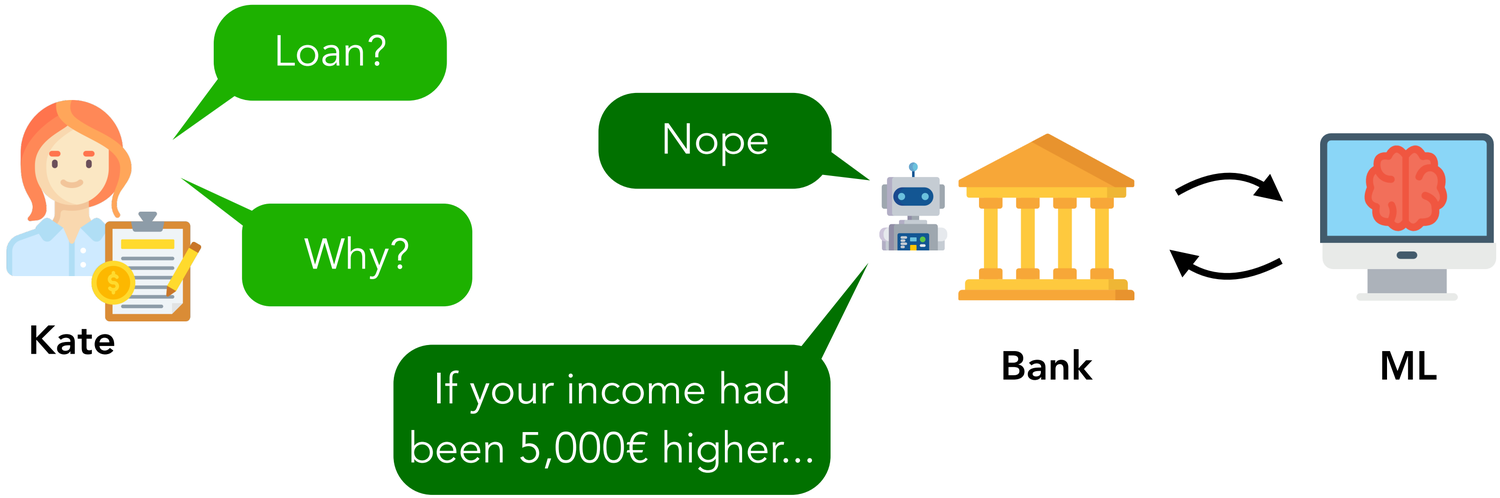
The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

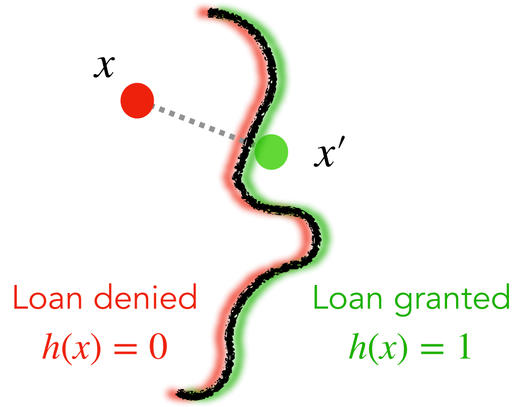
The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

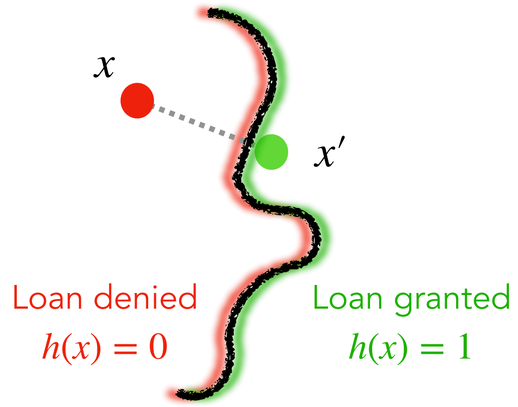
Given a (binary) prediction $h(x)$ by a machine learning model about an individual with features x, a counterfactual explanation is given by the closest feature value x' under which $h(x') \neq h(x)$



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

Given a (binary) prediction $h(x)$ by a machine learning model about an individual with features x, a counterfactual explanation is given by the closest feature value x' under which $h(x') \neq h(x)$



By showing a feature-perturbed version of an individual, a counterfactual explanations is, in principle, telling the individual what to do to secure a better decision in the future.

Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

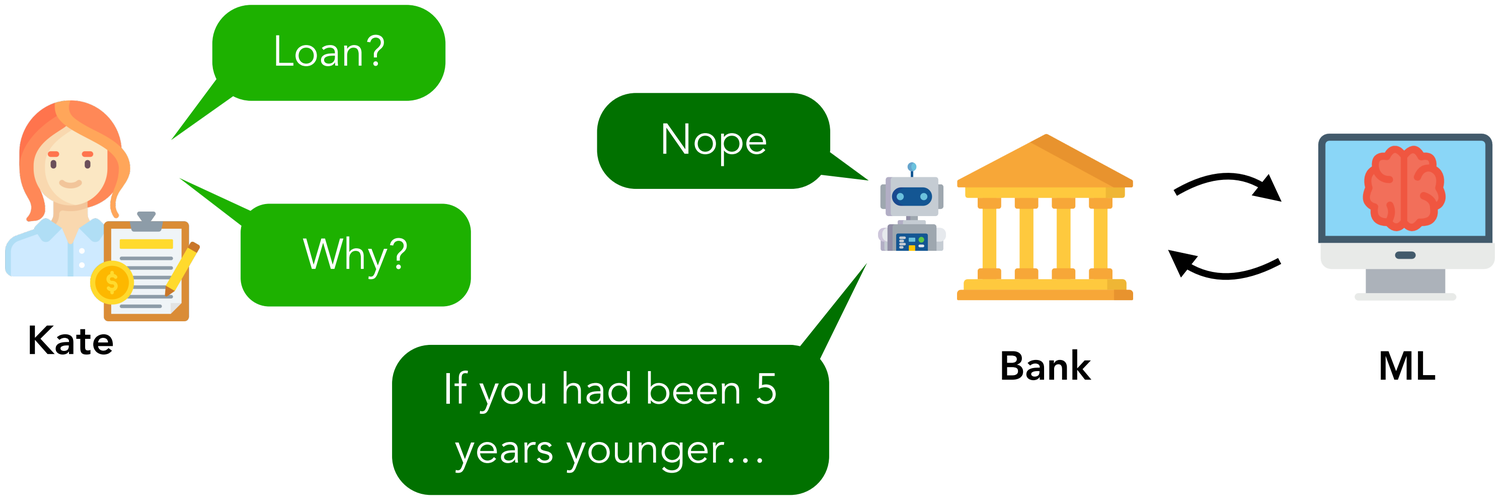
Counterfactual explanations

However, the closest feature value x' may not be actionable, and may not even be plausible.

Verma et al. "*Counterfactual explanations and algorithmic recourses for machine learning: A review.*" ACM Computing Surveys, 2024.

Counterfactual explanations

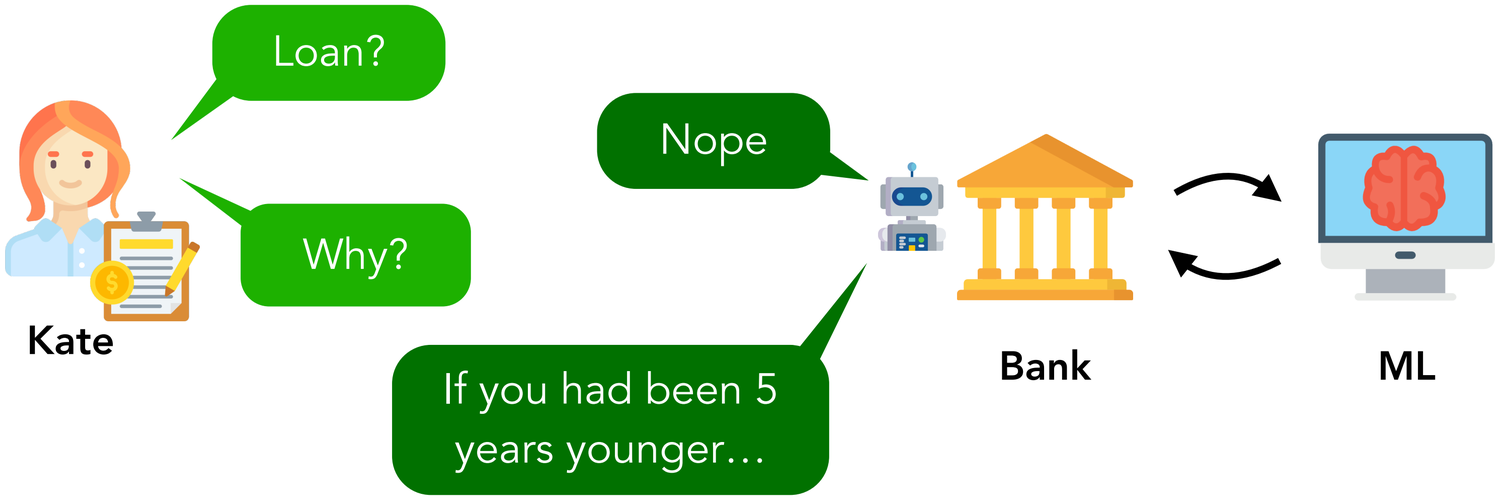
However, the closest feature value x' may not be actionable, and may not even be plausible.



Verma et al. "*Counterfactual explanations and algorithmic recourses for machine learning: A review.*" ACM Computing Surveys, 2024.

Counterfactual explanations

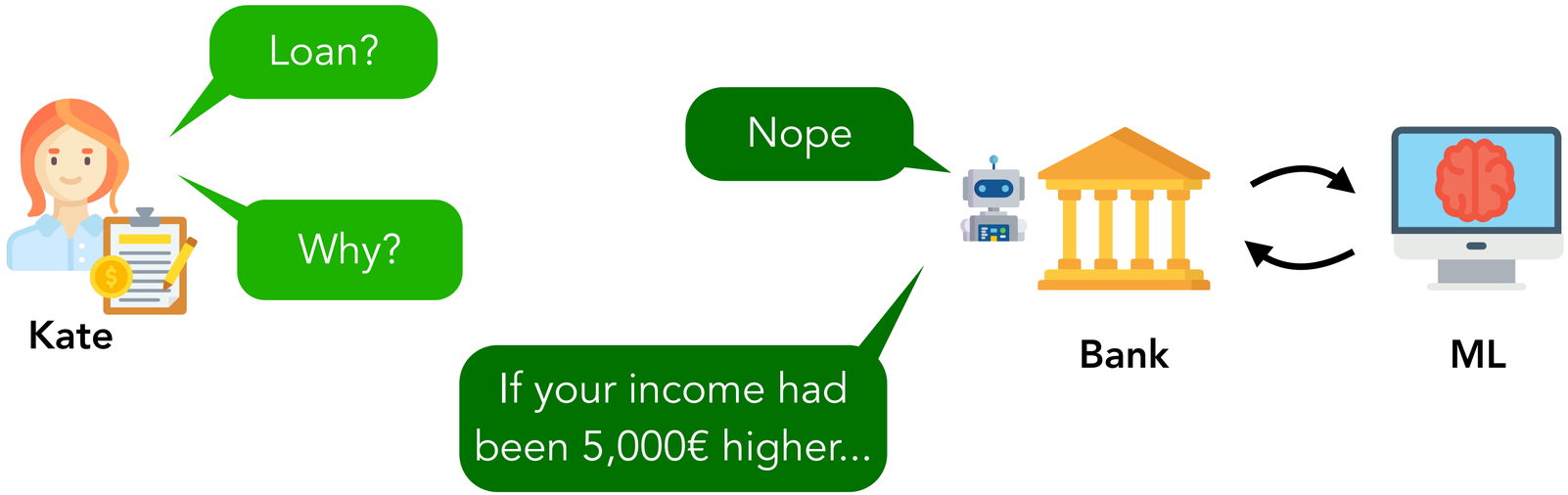
However, the closest feature value x' may not be actionable, and may not even be plausible.



Many follow-up works have addressed this problem by finding the closest feature value subject to a variety of actionability and plausibility constraints.

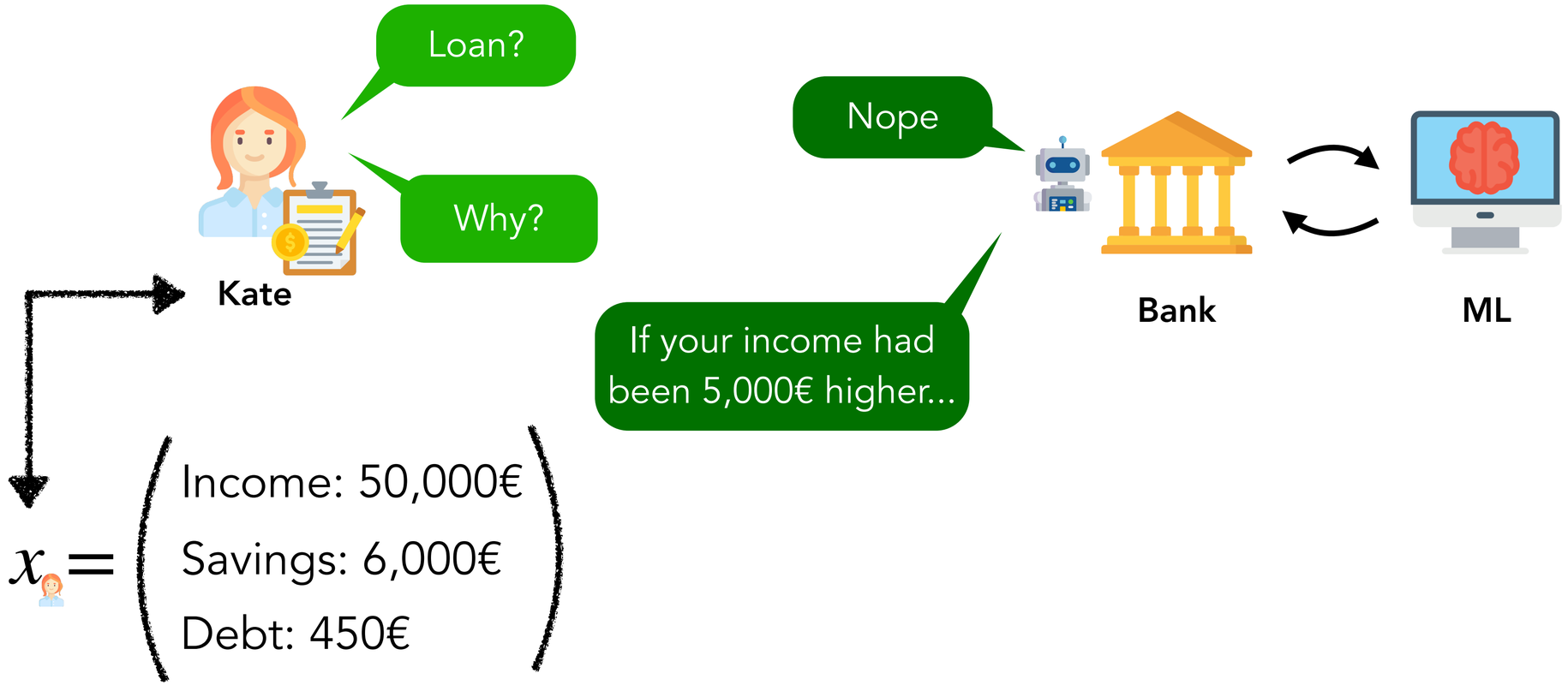
Verma et al. "*Counterfactual explanations and algorithmic recourses for machine learning: A review.*" ACM Computing Surveys, 2024.

Counterfactual explanations ignore causal dependencies



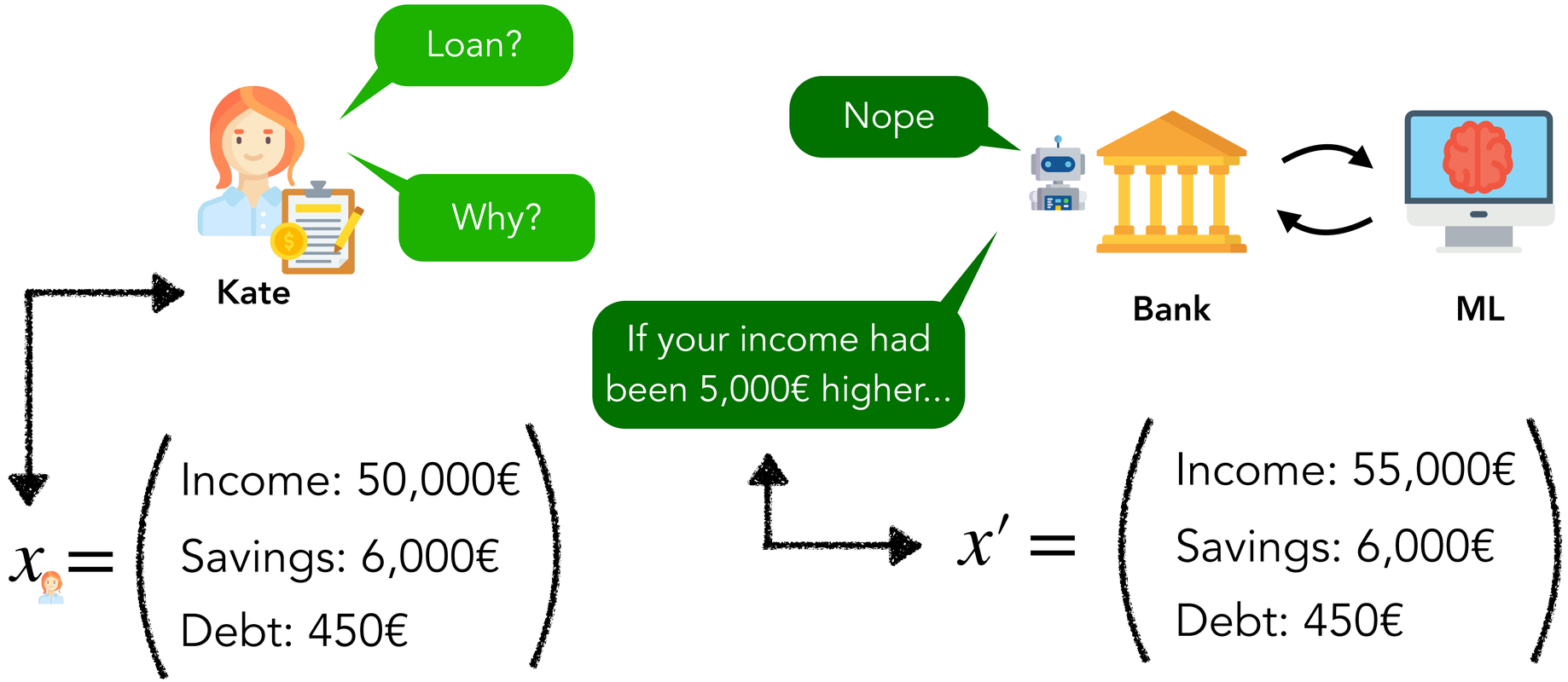
Beckers. "*Causal explanations and xai.*" CLeaR, 2022.
Crupi et al. "*Counterfactual explanations as interventions in latent space.*" DMKD, 2022.

Counterfactual explanations ignore causal dependencies



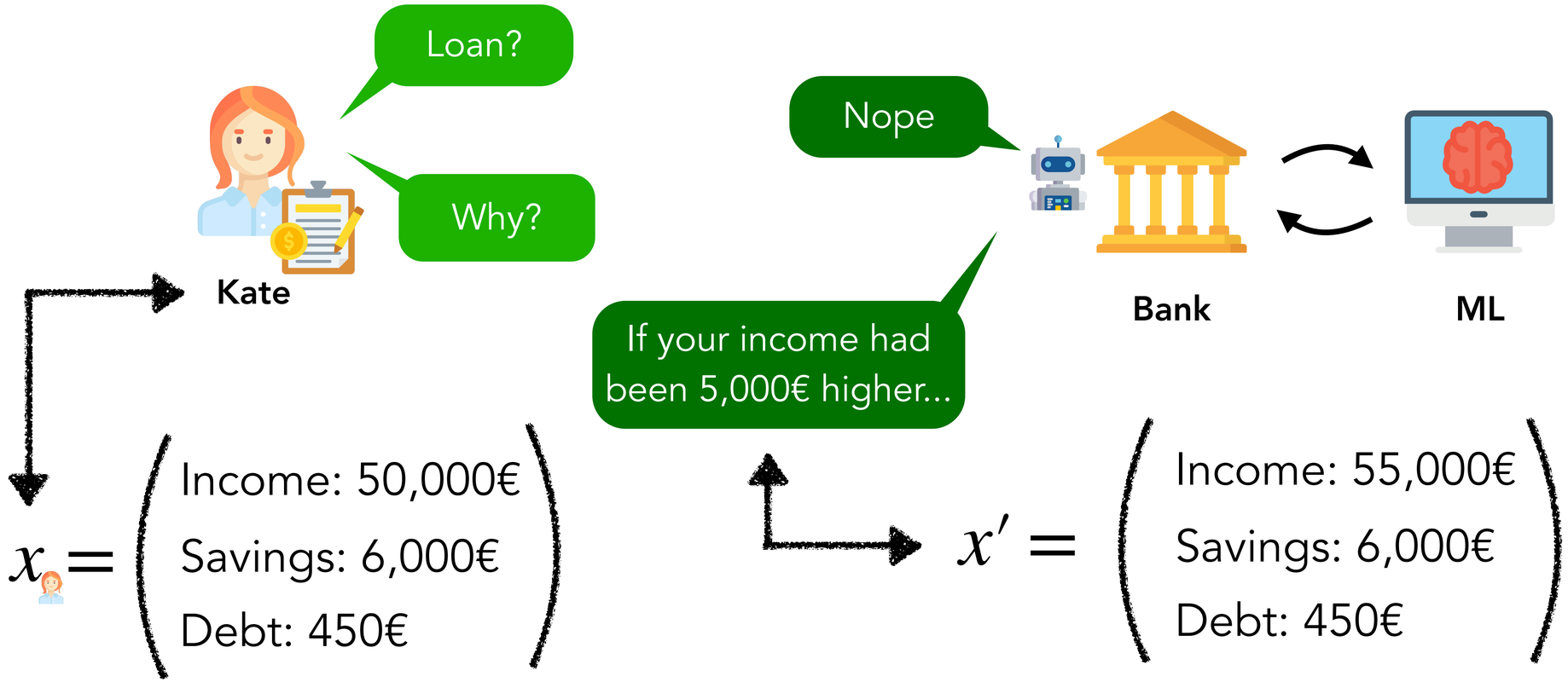
Beckers. "*Causal explanations and xai.*" CLeaR, 2022.
Crupi et al. "*Counterfactual explanations as interventions in latent space.*" DMKD, 2022.

Counterfactual explanations ignore causal dependencies



Beckers. "*Causal explanations and xai.*" CLeaR, 2022.
Crupi et al. "*Counterfactual explanations as interventions in latent space.*" DMKD, 2022.

Counterfactual explanations ignore causal dependencies



If Kate's income had been 5,000€ higher, Kate's savings would have been more than 6,000€!

Beckers. "*Causal explanations and xai.*" CLeaR, 2022.
Crupi et al. "*Counterfactual explanations as interventions in latent space.*" DMKD, 2022.

Counterfactual explanations as interventions

A counterfactual explanation does not answer a counterfactual question but an interventional question.

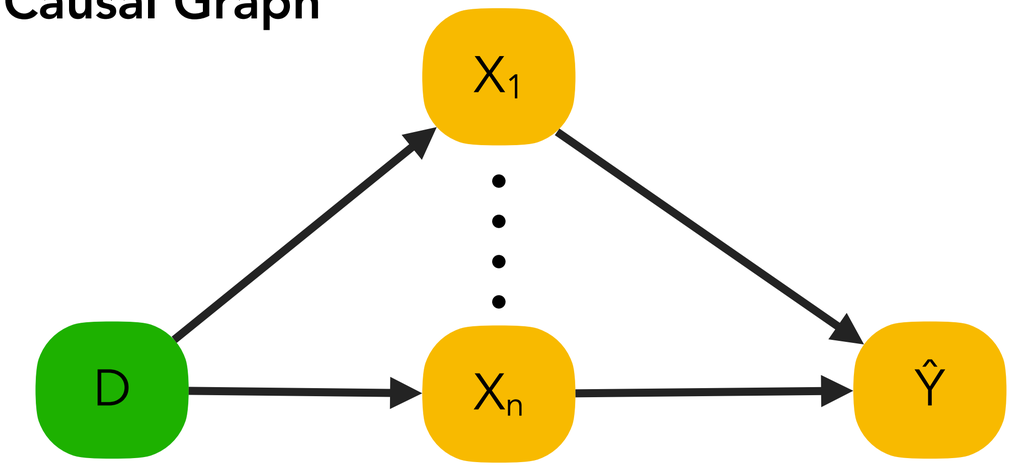
Counterfactual explanations as interventions

A counterfactual explanation does not answer a counterfactual question but an interventional question.

Structural Causal Model $\mathscr{M}$

$$X_1 := f_{X_1}(D)$$
$$\vdots$$
$$X_n := f_{X_n}(D)$$
$$\hat{Y} := h(X)$$
$$D \sim P(D)$$

Causal Graph



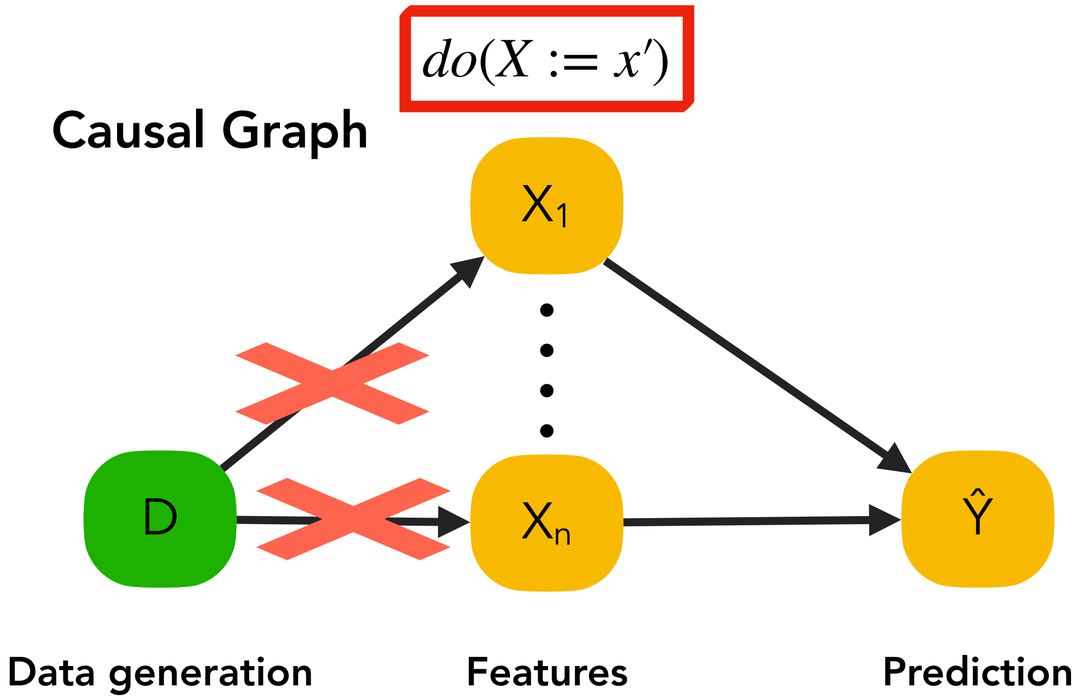
Counterfactual explanations as interventions

A counterfactual explanation does not answer a counterfactual question but an interventional question.

Structural Causal Model $\mathscr{M}$

$$X_1 := f_1(D) \quad X_1 := x_1'$$
$$\vdots \qquad\qquad \vdots$$
$$X_n := f_n(D) \quad X_n := x_n'$$
$$\hat{Y} := h(X)$$
$$D \sim P(D)$$

Causal Graph



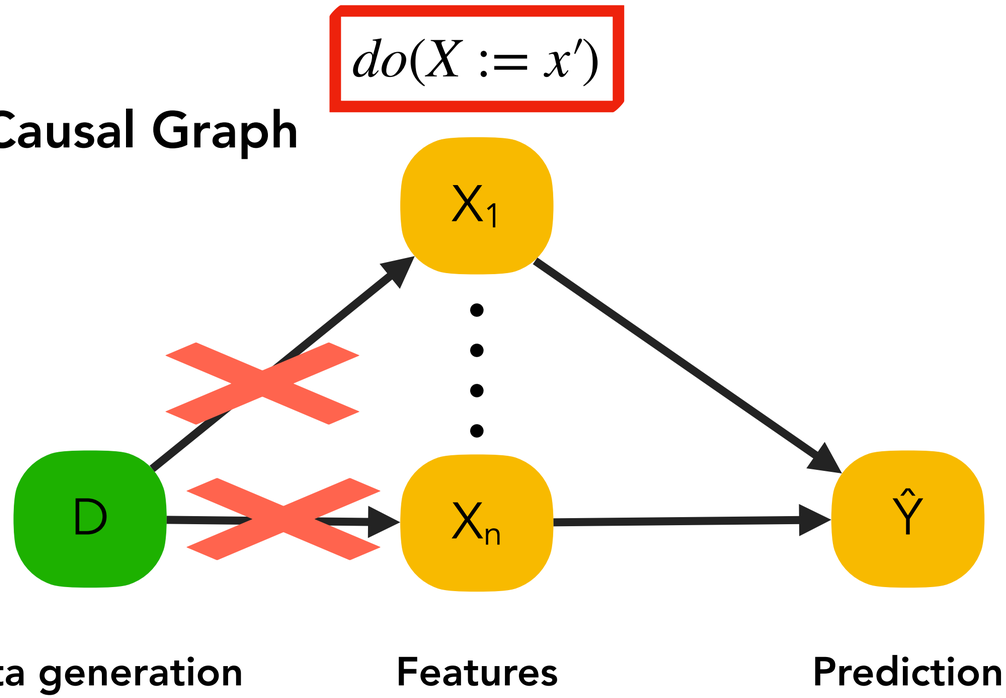
Counterfactual explanations as interventions

A counterfactual explanation does not answer a counterfactual question but an interventional question.

Structural Causal Model $\mathscr{M}$

$$X_1 := f_1(D) \quad X_1 := x_1'$$
$$\vdots \qquad\qquad \vdots$$
$$X_n := f_n(D) \quad X_n := x_n'$$
$$\hat{Y} := h(X)$$
$$D \sim P(D)$$

Causal Graph



A counterfactual explanation encourages an individual to change the value of the features x_l such that $x_l \neq x_l'$. However, it does not take into account that such a change may induce changes in features x_l such that $x_l = x_l'$.

Algorithmic recourse

Algorithmic recourse seeks to find the minimal intervention a under which $h(x + a) \neq h(x)$ *while accounting for causal dependencies between features.*

Karimi et al. "*Algorithmic recourse: from counterfactual explanations to interventions.*" FAccT, 2021.
Karimi et al. "*Algorithmic recourse under imperfect causal knowledge: a probabilistic approach.*" NeurIPS, 2020.

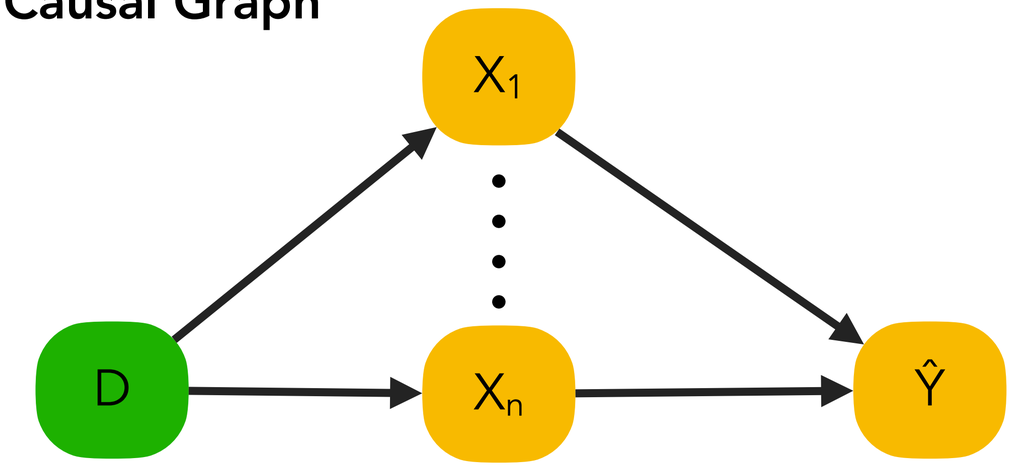
Algorithmic recourse

Algorithmic recourse seeks to find the minimal intervention a under which $h(x+a) \neq h(x)$ *while accounting for causal dependencies between features.*

Structural Causal Model $\mathscr{M}$

$$X_1 := f_{X_1}(D)$$
$$\vdots$$
$$X_n := f_{X_n}(D)$$
$$\hat{Y} := h(X)$$
$$D \sim P(D)$$

Causal Graph



Karimi et al. "*Algorithmic recourse: from counterfactual explanations to interventions.*" FAccT, 2021.
Karimi et al. "*Algorithmic recourse under imperfect causal knowledge: a probabilistic approach.*" NeurIPS, 2020.

Algorithmic recourse

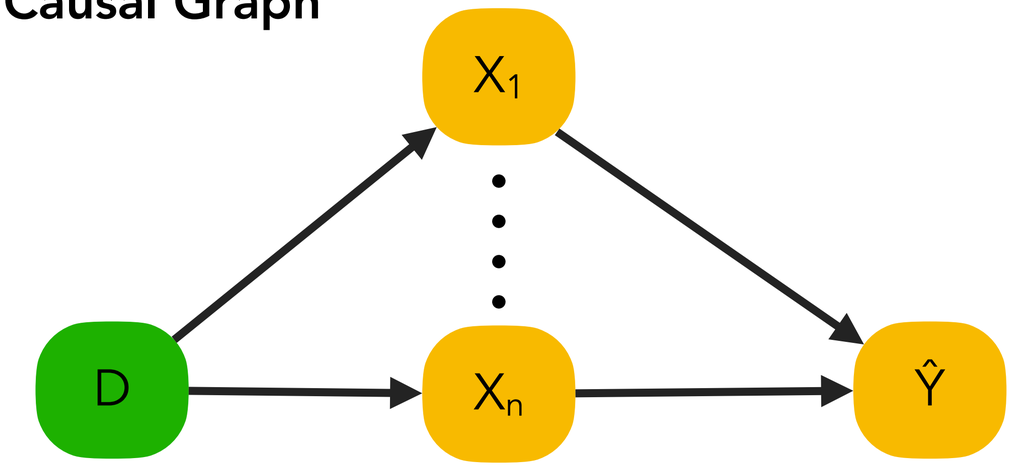
Algorithmic recourse seeks to find the minimal intervention a under which $h(x+a) \neq h(x)$ *while accounting for causal dependencies between features.*

Modified Structural Causal Model $\mathscr{M}_{X=x}$

$$X_1 := f_{X_1}(D)$$
$$\vdots$$
$$X_n := f_{X_n}(D)$$
$$\hat{Y} := h(X)$$
$$D \sim P(D \mid X = x)$$

Posterior distribution of the noise

Causal Graph



Karimi et al. "*Algorithmic recourse: from counterfactual explanations to interventions.*" FAccT, 2021.
Karimi et al. "*Algorithmic recourse under imperfect causal knowledge: a probabilistic approach.*" NeurIPS, 2020.

Algorithmic recourse

Algorithmic recourse seeks to find the minimal intervention a under which $h(x+a) \neq h(x)$ *while accounting for causal dependencies between features.*

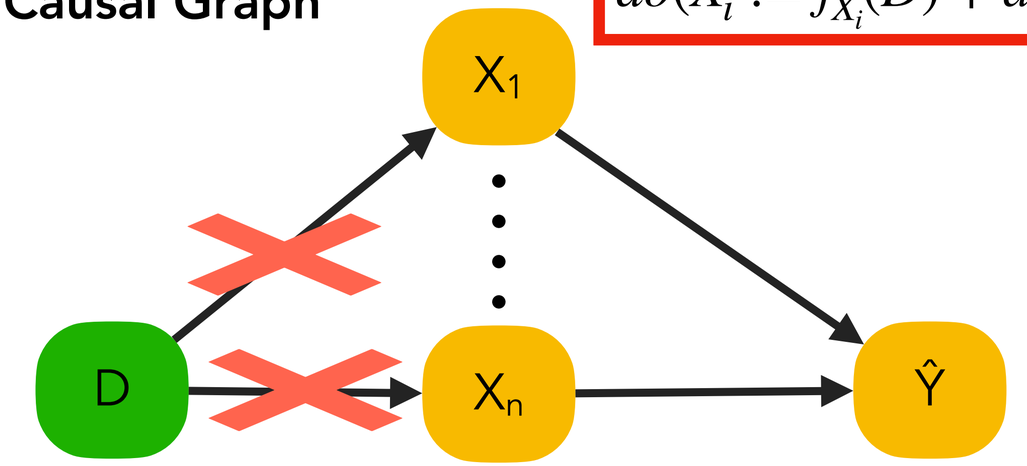
Modified Structural Causal Model $\mathscr{M}_{X=x}$

$$X_1 := f_{X_1}(D) + a_1$$
$$\vdots$$
$$X_n := f_{X_n}(D) + a_n$$
$$\hat{Y} := h(X)$$
$$D \sim P(D \mid X = x)$$

Posterior distribution of the noise

Causal Graph

$do(X_i := f_{X_i}(D) + a_i)$



Whenever $a_i = 0$, the value of X_i may still change!

Karimi et al. "*Algorithmic recourse: from counterfactual explanations to interventions.*" FAccT, 2021.
Karimi et al. "*Algorithmic recourse under imperfect causal knowledge: a probabilistic approach.*" NeurIPS, 2020.

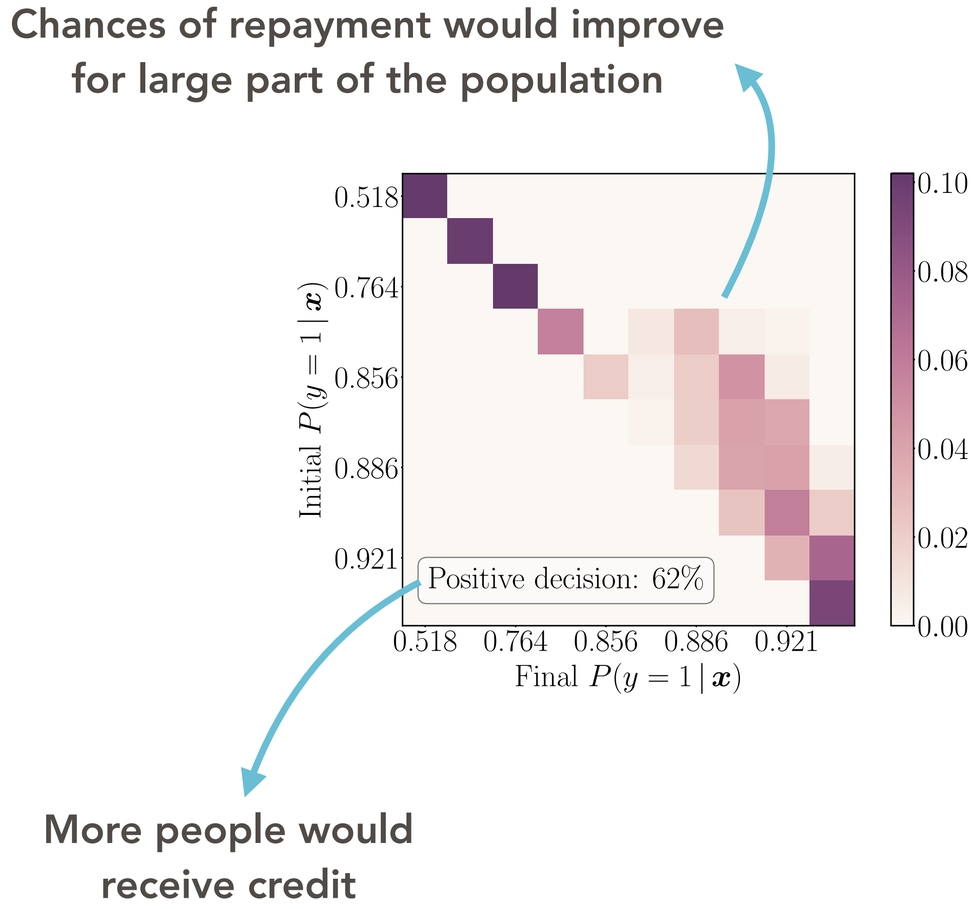
Counterfactual explanations & performativity

If a sizable number of individuals follow the changes prescribed by counterfactual explanations, the feature distribution $P(X)$ may change.

Tsirtsis and Gomez-Rodriguez. "*Decisions, counterfactual explanations and strategic behavior.*" NeurIPS, 2020.
Perdomo et al. "*Performative prediction.*" ICML, 2020.

Counterfactual explanations & performativity

If a sizable number of individuals follow the changes prescribed by counterfactual explanations, the feature distribution $P(X)$ may change.



Tsirtsis and Gomez-Rodriguez. "*Decisions, counterfactual explanations and strategic behavior.*" NeurIPS, 2020.
Perdomo et al. "*Performative prediction.*" ICML, 2020.

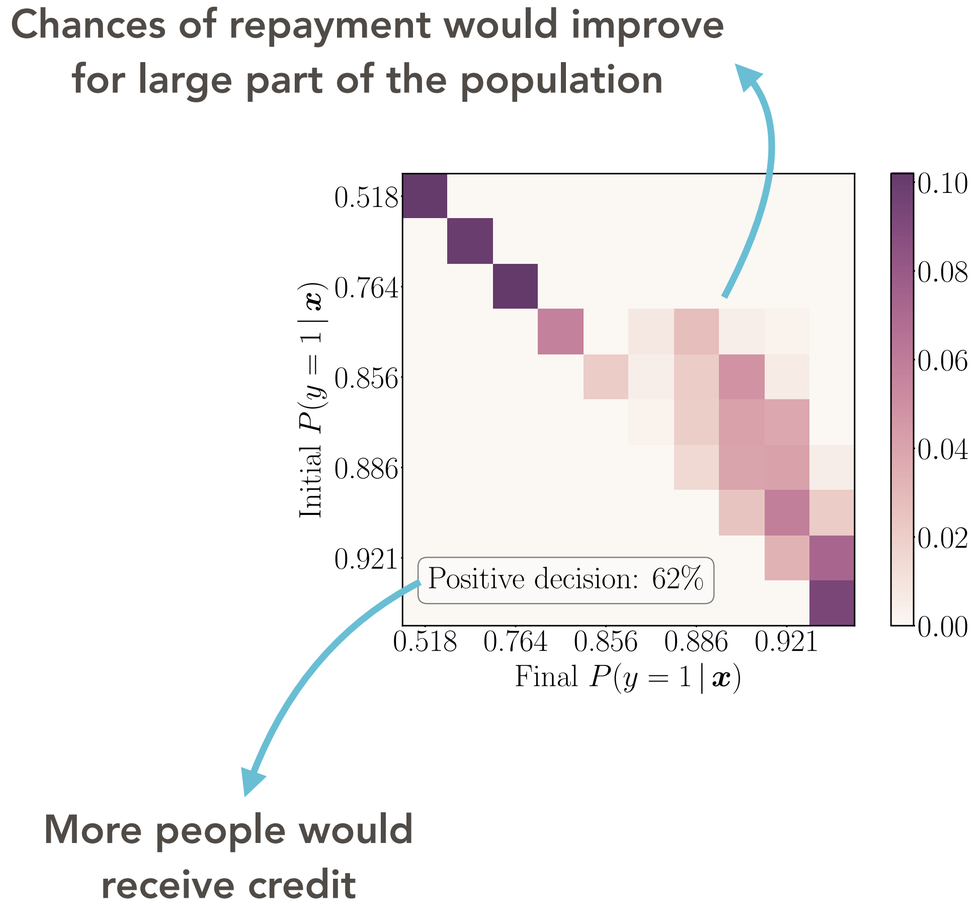
Counterfactual explanations & performativity

If a sizable number of individuals follow the changes prescribed by counterfactual explanations, the feature distribution $P(X)$ may change.

This raises the question of finding decision policies π and counterfactual explanations $\mathscr{A}$ that are optimal in terms of utility.

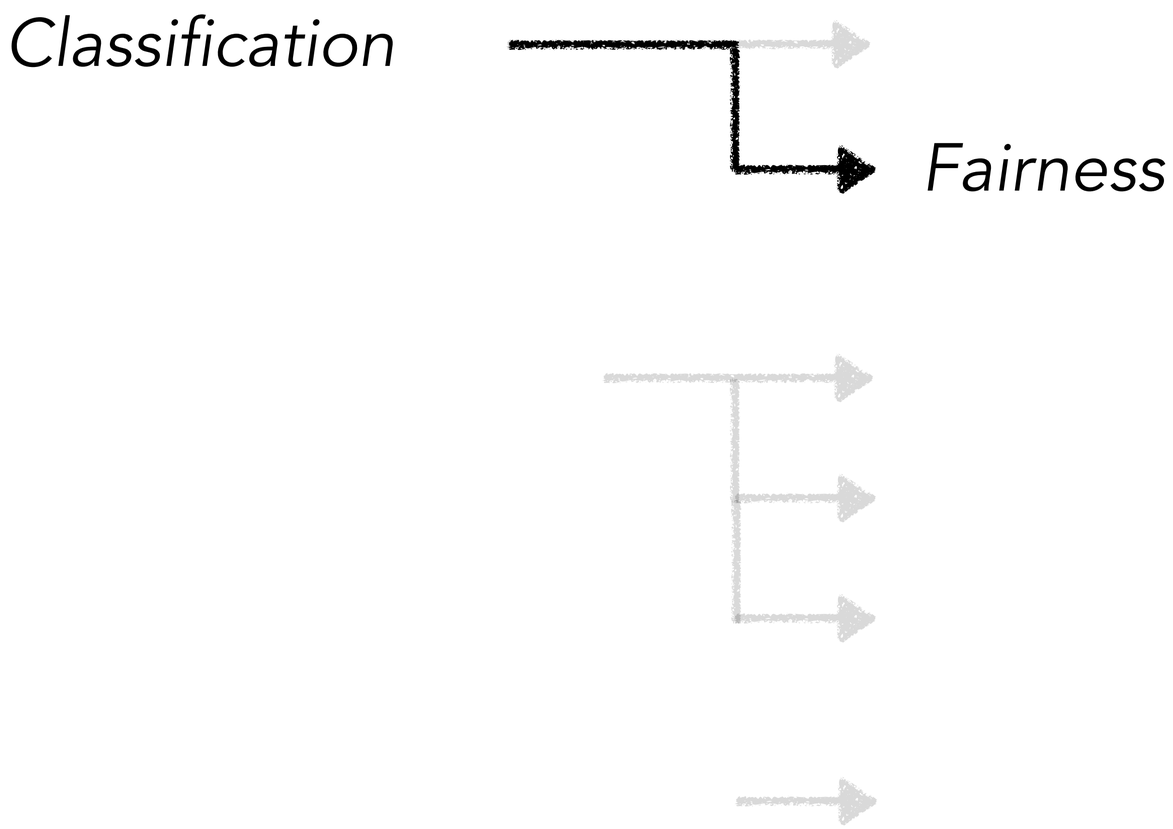
$$\max_{\pi,\mathscr{A}} u(\pi, \mathscr{A}) := \mathbb{E}_{x \sim P(X \mid \pi, \mathscr{A})} \Big[\pi(x) \big(P(Y = 1 \mid x) - \gamma \big) \Big]$$

constant reflecting economic considerations of the decision maker



Tsirtsis and Gomez-Rodriguez. "*Decisions, counterfactual explanations and strategic behavior.*" NeurIPS, 2020.
Perdomo et al. "*Performative prediction.*" ICML, 2020.

Use cases of counterfactuals in machine learning



Counterfactual fairness

Counterfactual fairness captures the intuition that a prediction by a machine learning model is fair towards an individual who belongs to a demographic group $A = a$ if it would have been the same had the individual belonged to a different demographic group $A = a'$.

Kusner et al. "*Counterfactual fairness.*" NeurIPS, 2017.

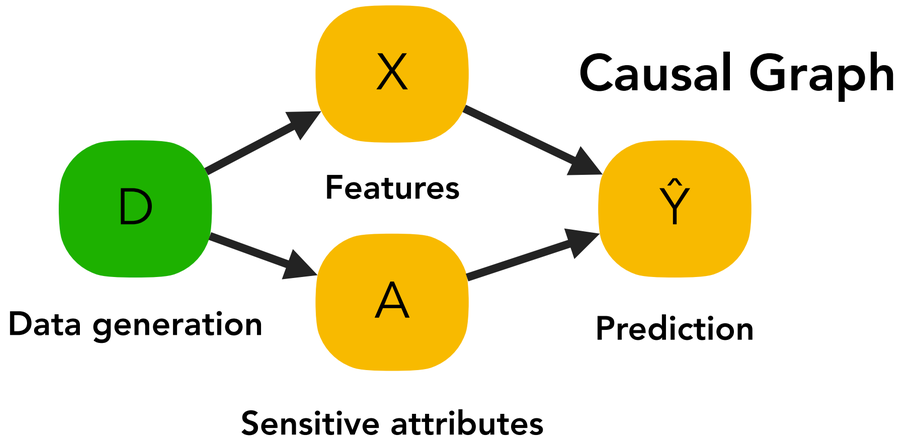
Counterfactual fairness

Counterfactual fairness captures the intuition that a prediction by a machine learning model is fair towards an individual who belongs to a demographic group $A = a$ if it would have been the same had the individual belonged to a different demographic group $A = a'$.

Structural Causal Model $\mathscr{M}$

$$X := f_X(D) \qquad \hat{Y} := h(X, A)$$

$$A := f_A(D) \qquad D \sim P(D)$$



Kusner et al. "*Counterfactual fairness.*" NeurIPS, 2017.

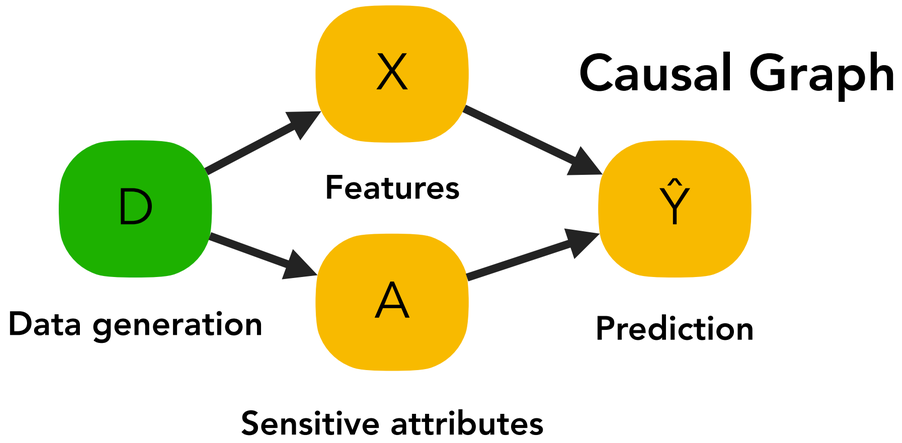
Counterfactual fairness

Counterfactual fairness captures the intuition that a prediction by a machine learning model is fair towards an individual who belongs to a demographic group $A = a$ if it would have been the same had the individual belonged to a different demographic group $A = a'$

Structural Causal Model $\mathscr{M}$

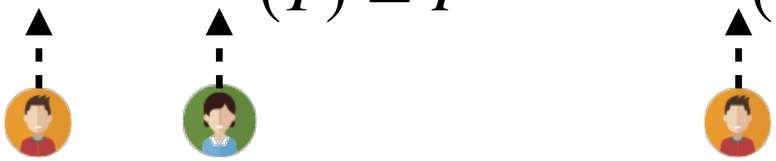
$$X := f_X(D) \qquad \hat{Y} := h(X, A)$$

$$A := f_A(D) \qquad D \sim P(D)$$



Counterfactual fairness

$$P^{\mathscr{M}\,|\,X=x, A=a\,;\,do(A=a')}(\hat{Y}) = P^{\mathscr{M}\,|\,X=x, A=a}(\hat{Y})$$



Kusner et al. "*Counterfactual fairness.*" NeurIPS, 2017.

Counterfactual fairness can be too restrictive

Counterfactual fairness considers the full effect of the demographic group on the prediction as problematic. However, this is not the case in certain scenarios.

Chiappa. "*Path-specific counterfactual fairness.*" AAAI, 2019.

Counterfactual fairness can be too restrictive

Counterfactual fairness considers the full effect of the demographic group on the prediction as problematic. However, this is not the case in certain scenarios.

Alleged gender bias case at Berkeley


8,442 male applicants for the fall of 1973, 44 percent were admitted,


4,351 female applicants, 35 percent were admitted

Chiappa. "*Path-specific counterfactual fairness.*" AAAI, 2019.

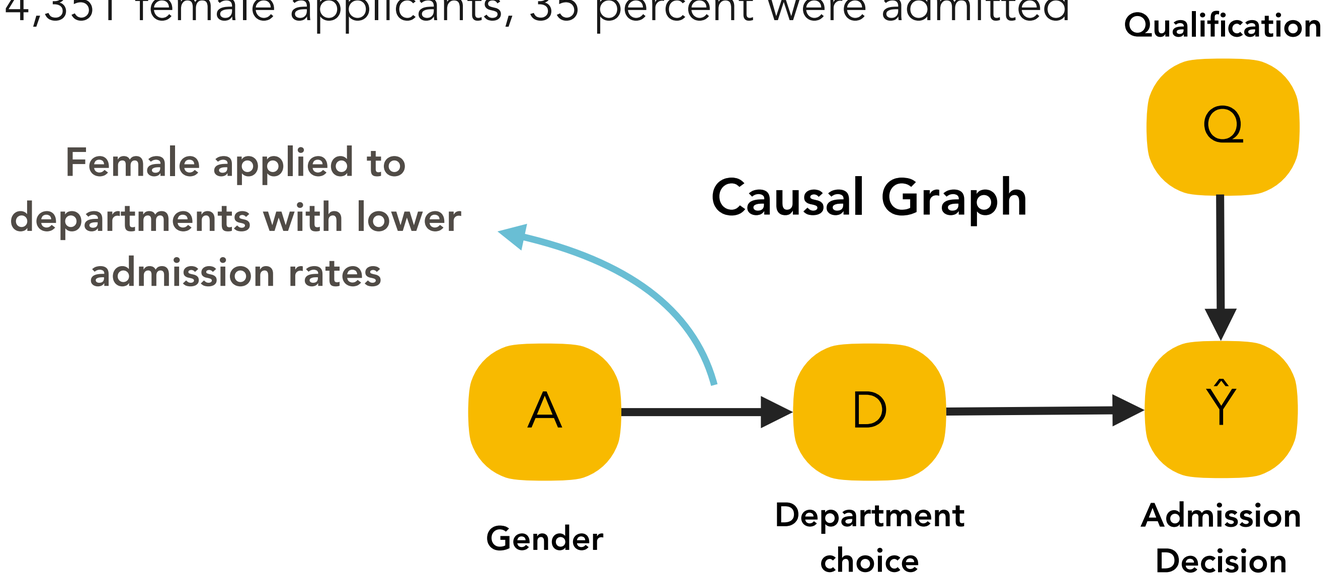
Counterfactual fairness can be too restrictive

Counterfactual fairness considers the full effect of the demographic group on the prediction as problematic. However, this is not the case in certain scenarios.

Alleged gender bias case at Berkeley


8,442 male applicants for the fall of 1973, 44 percent were admitted,


4,351 female applicants, 35 percent were admitted



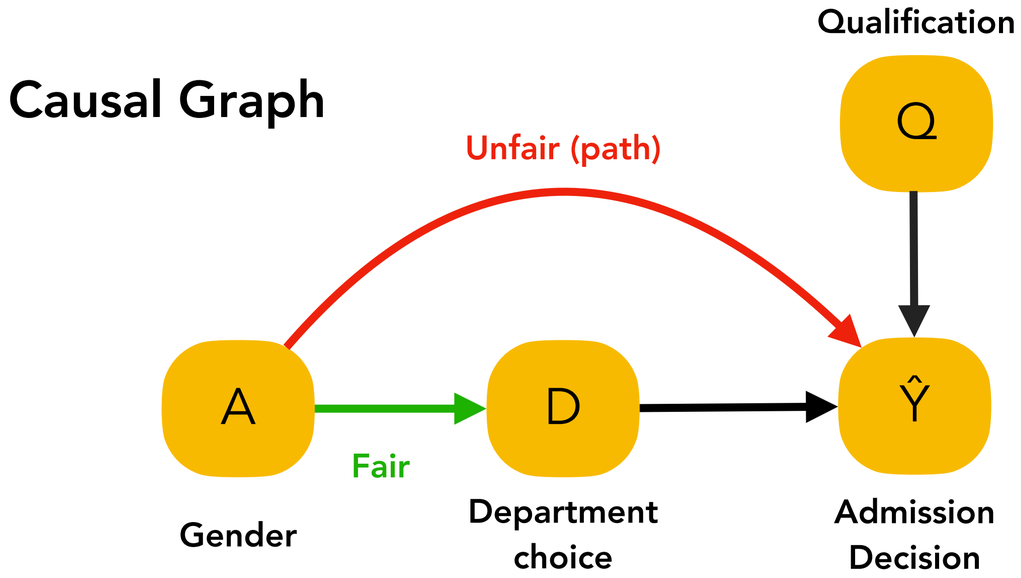
Counterfactual fairness is violated

$$P^{\mathcal{M}}(\hat{Y}) \neq P^{\mathcal{M}}(\hat{Y})$$

Chiappa. "*Path-specific counterfactual fairness.*" AAAI, 2019.

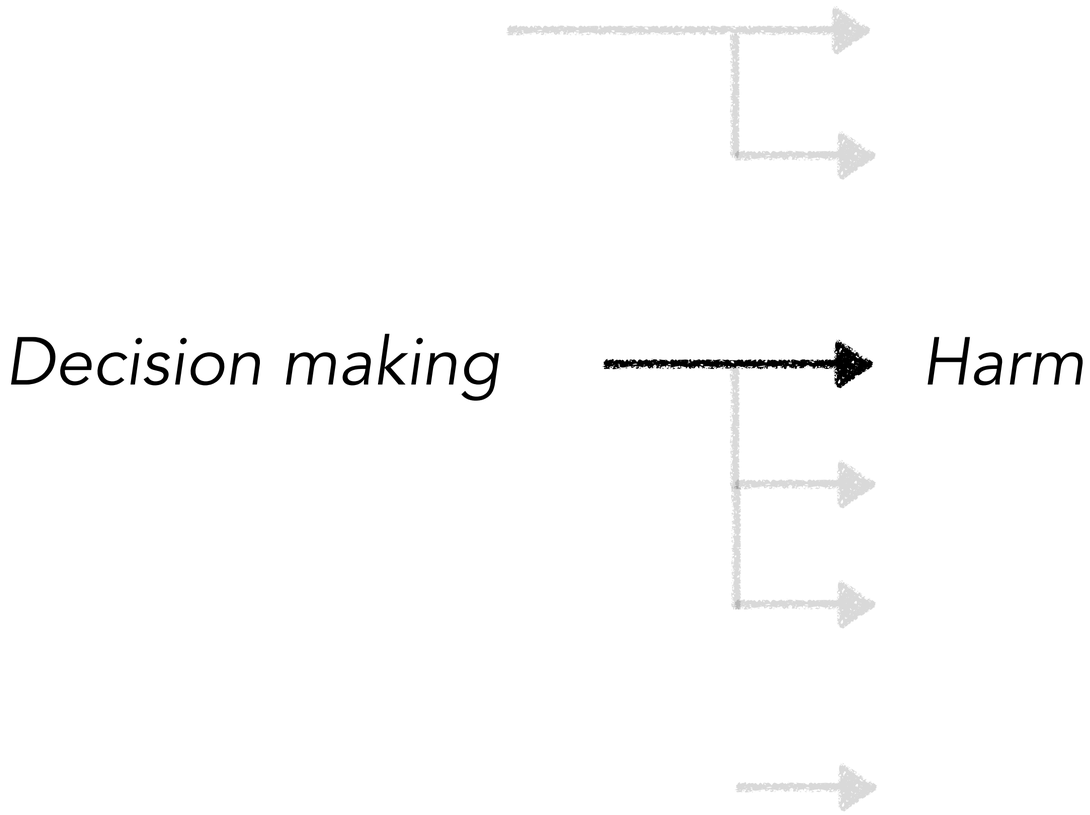
Path-specific counterfactual fairness

Path-specific counterfactual fairness is a more fine-grained fairness criterion that deals with sensitive attributes affecting the prediction along both fair and unfair pathways.

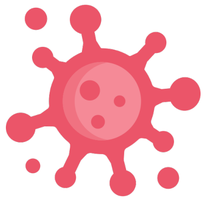


Chiappa. "*Path-specific counterfactual fairness.*" AAAI, 2019.

Use cases of counterfactuals in machine learning



Counterfactual harm



Disease

50% mortality rate



Treatment A

60% chance of curing a patient

40% chance of having no effect

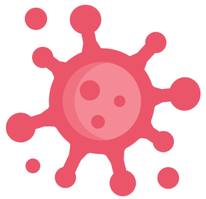


Treatment B

80% chance of curing a patient

20% chance of killing a patient

Counterfactual harm



Disease

50% mortality rate



Treatment A

60% chance of curing a patient

40% chance of having no effect



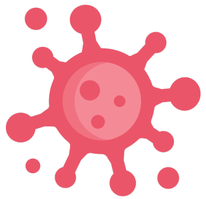
Treatment B

80% chance of curing a patient

20% chance of killing a patient

Treatments A and B have **identical recovery rates**. However, doctors would systematically favor treatment A as it achieves the same recovery rate but never harms the patient.

Counterfactual harm



Disease
50% mortality rate



Treatment A

60% chance of curing a patient

40% chance of having no effect



Treatment B

80% chance of curing a patient

20% chance of killing a patient

Treatments A and B have **identical recovery rates**. However, doctors would systematically favor treatment A as it achieves the same recovery rate but never harms the patient.

- Under treatment A, there are no patients that would have survived had they not been treated.
- Under treatment B, there are patients who die following treatment who would have lived had they not been treated.

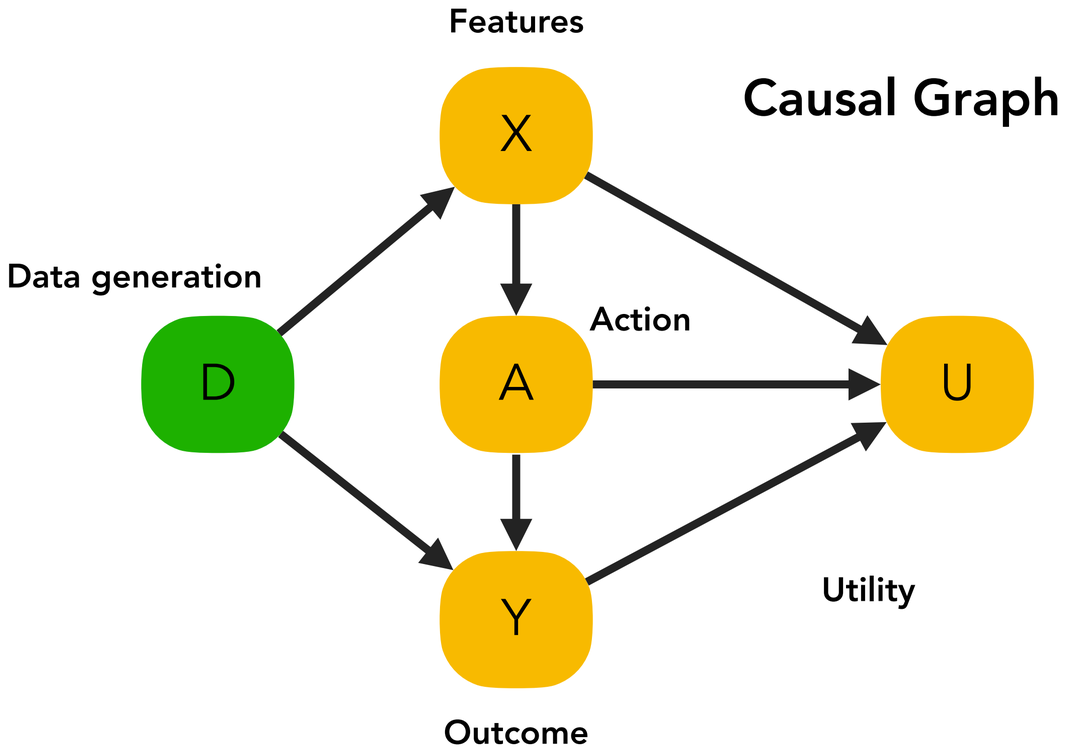
Formalizing counterfactual harm

Structural Causal Model $\mathscr{M}$

$X := f_X(D) \qquad Y := f_Y(D) \qquad D \sim P(D)$

$A := \pi(X)$ ◄- - **Algorithmic policy** 

$U := f_U(A, X, Y)$



Richens et al. "*Counterfactual harm.*" NeurIPS, 2022.

Formalizing counterfactual harm

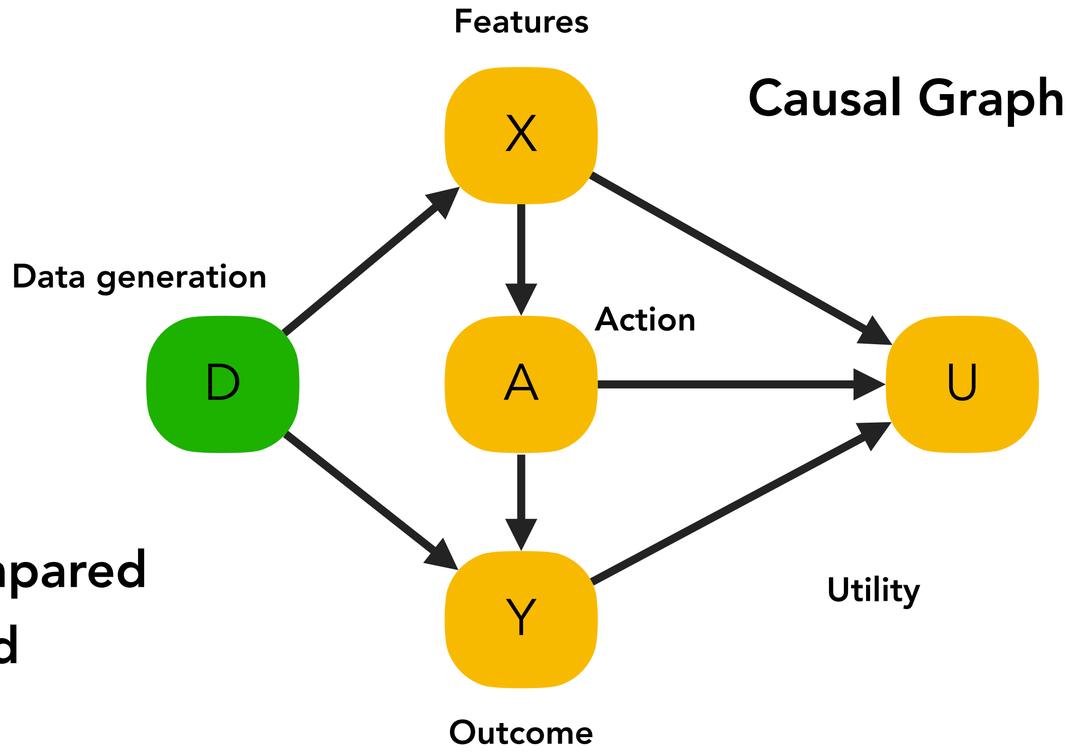
Structural Causal Model $\mathscr{M}$

$$X := f_X(D) \qquad Y := f_Y(D) \qquad D \sim P(D)$$

$$A := \pi(X)$$ ◄- - **Algorithmic policy**

$$U := f_U(A, X, Y)$$

Harm caused by action a taken by **compared to default action $\bar{a}$ given context $X = x$ and outcome $Y = y$**



Richens et al. "*Counterfactual harm.*" NeurIPS, 2022.

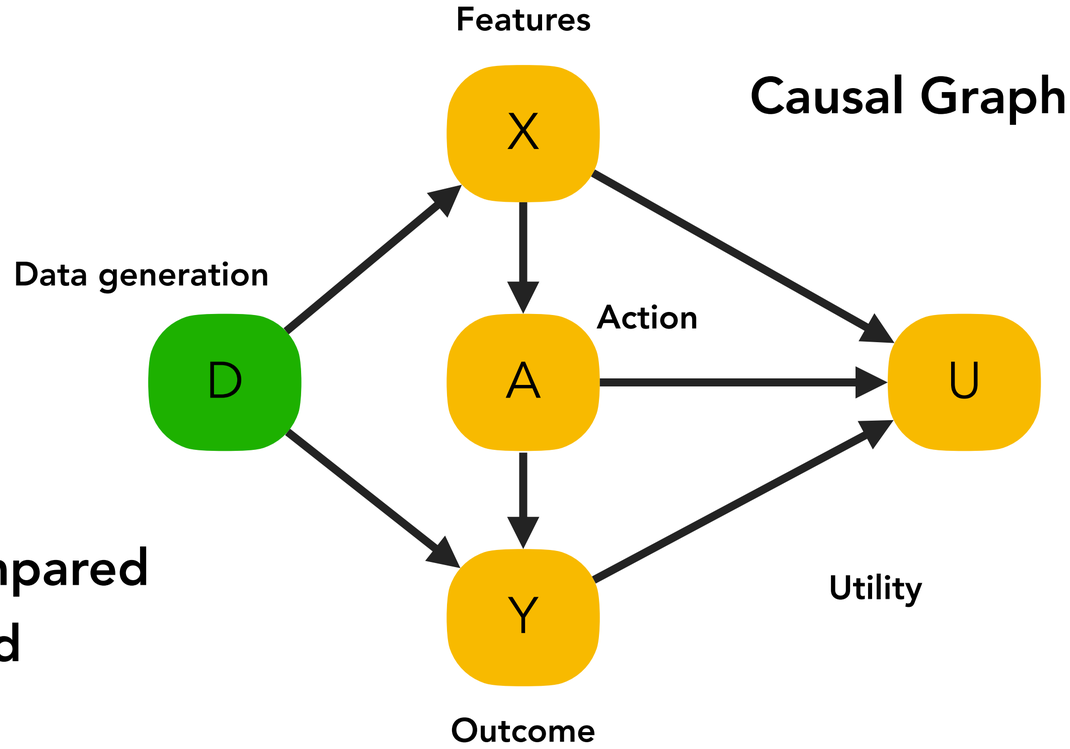
Formalizing counterfactual harm

Structural Causal Model $\mathscr{M}$

$$X := f_X(D) \qquad Y := f_Y(D) \qquad D \sim P(D)$$

$$A := \pi(X)$$ ◄- - **Algorithmic policy**

$$U := f_U(A, X, Y)$$



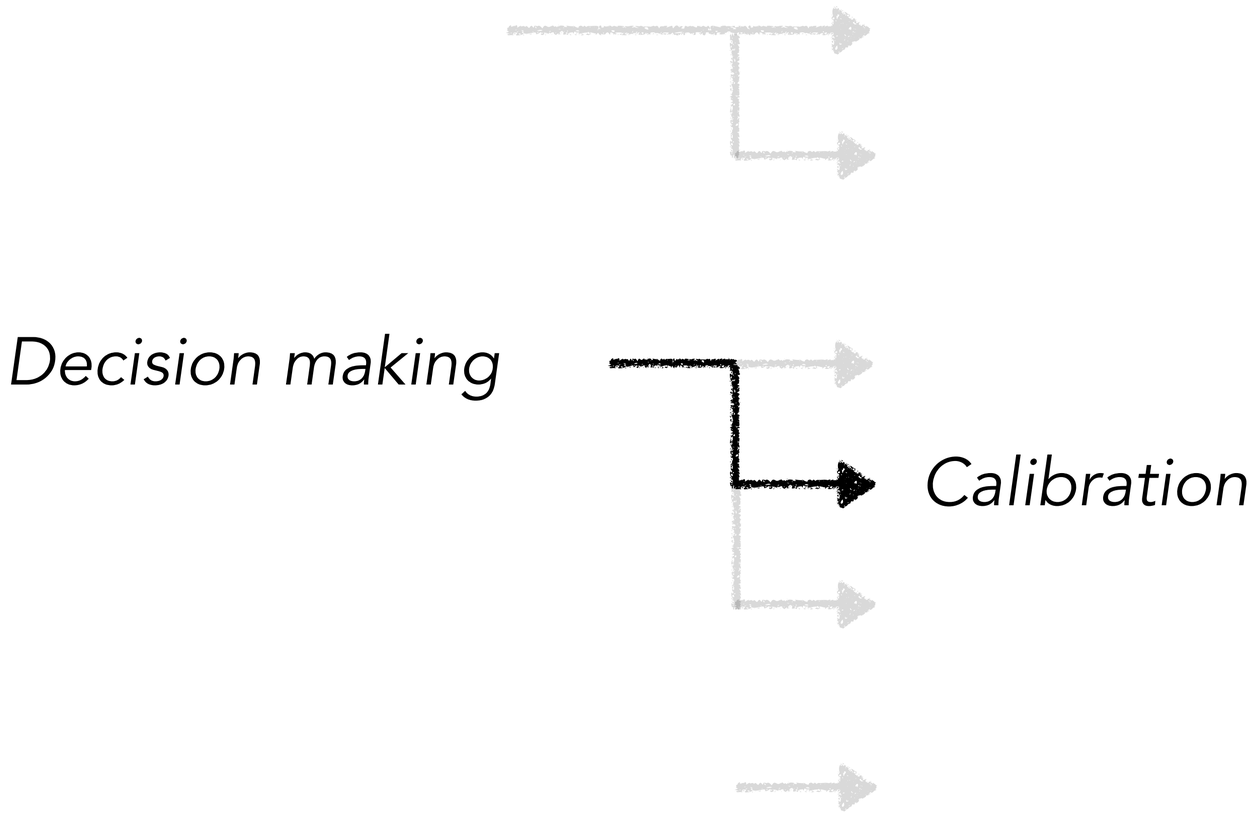
Harm caused by action a taken by **compared to default action $\bar{a}$ given context $X = x$ and outcome $Y = y$**

$$h(a, x, y) = \int_{y'} P^{\mathscr{M} \mid X=x,\, Y=y,\, A=a\,;\, do(A=\bar{a})}(Y = y') \max\left(0,\, U(\bar{a}, x, y') - U(a, x, y)\right) dy'$$

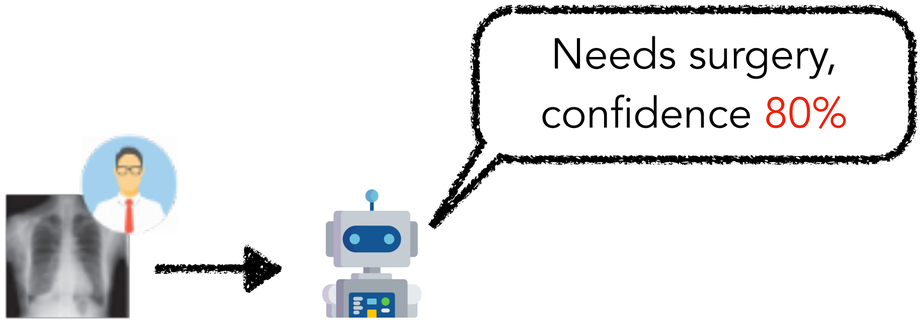
Counterfactual utility Utility

Richens et al. "*Counterfactual harm.*" NeurIPS, 2022.

Use cases of counterfactuals in machine learning



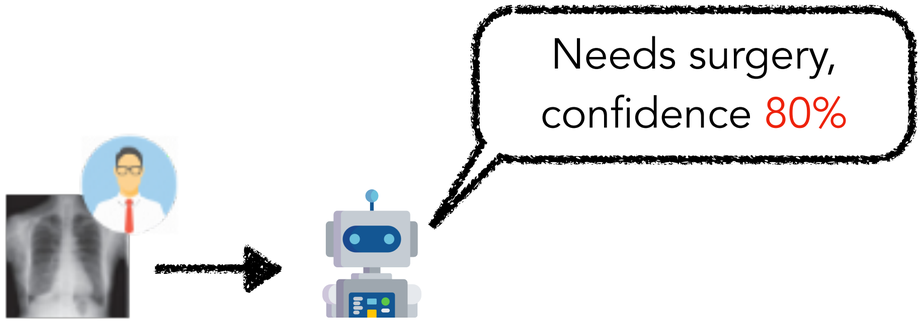
Calibration



Calibration:
Across all patients who predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration

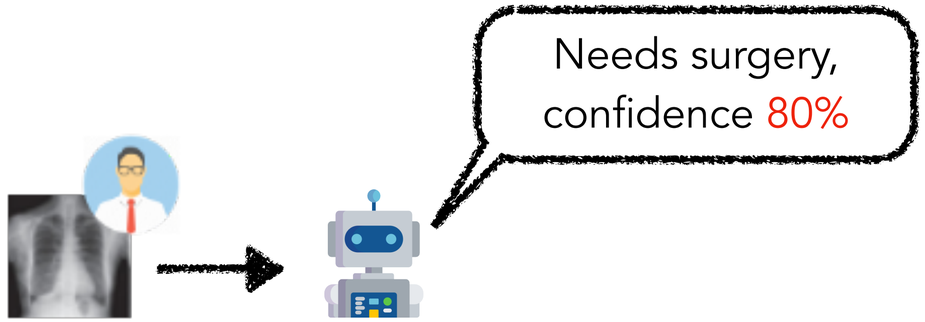


Calibration:
Across all patients who predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.

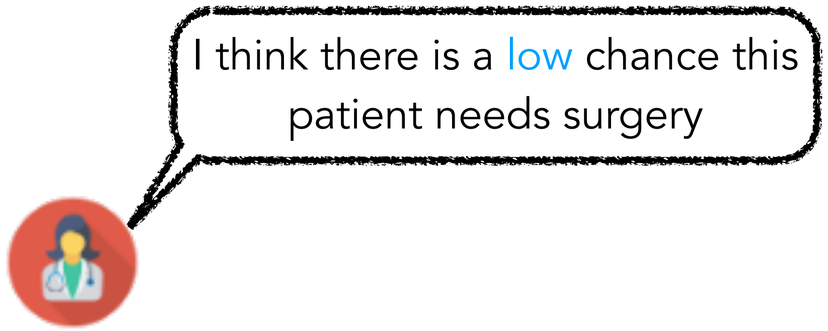
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



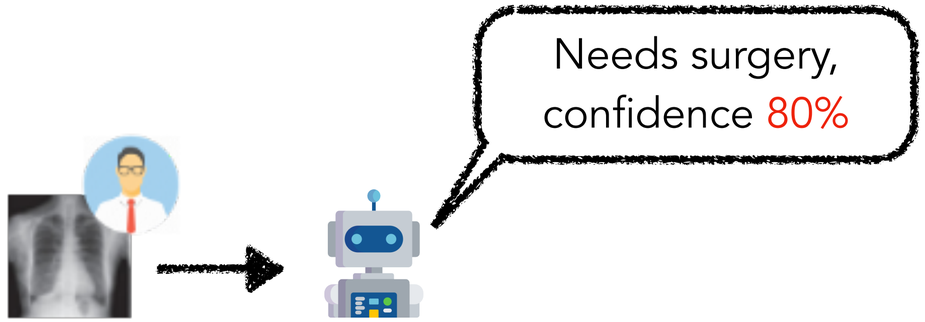
Calibration:
Across all patients who predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.



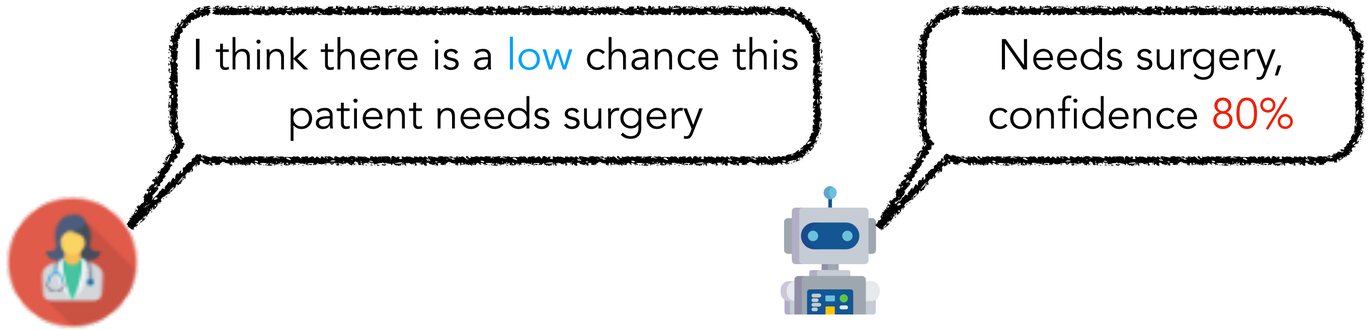
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



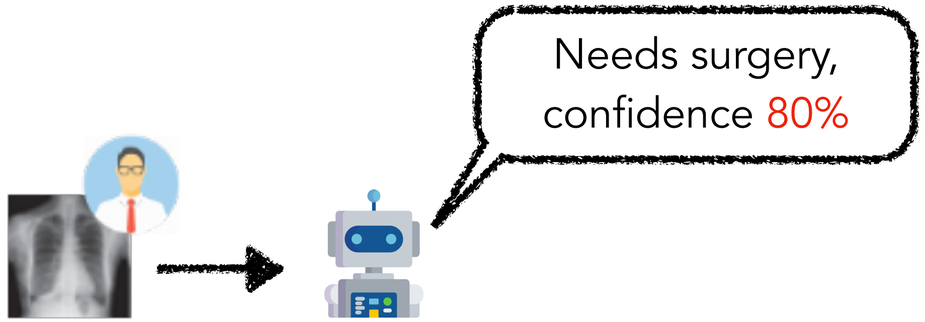
Calibration:
Across all patients who predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.



Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



Calibration:
Across all patients who predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

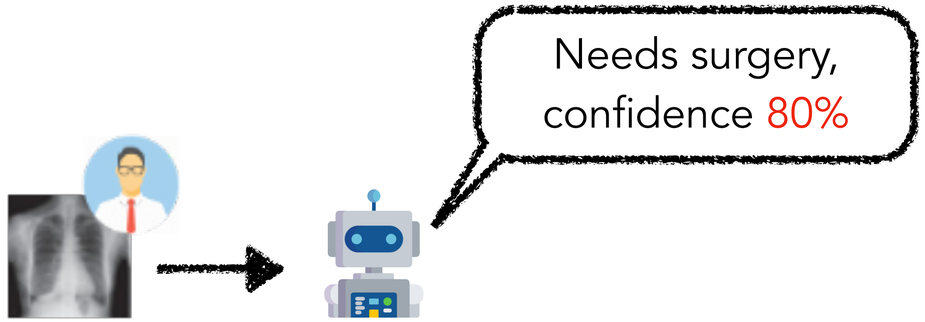
Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.



The doctor decides *optimally*

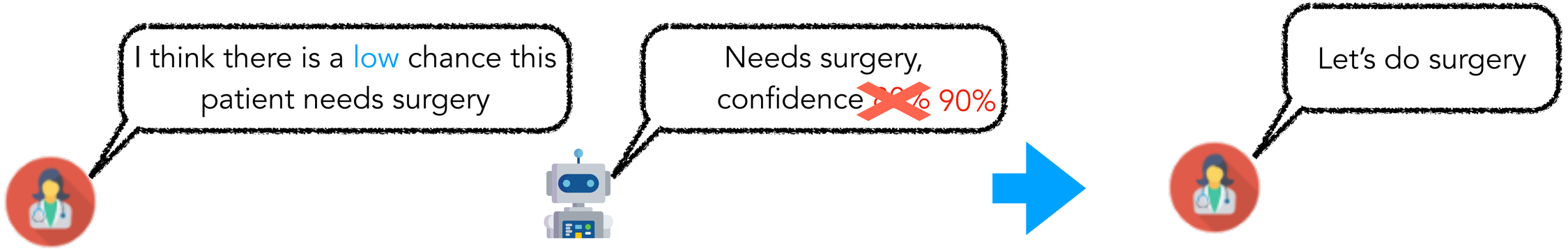
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



Calibration:
Across all patients who predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

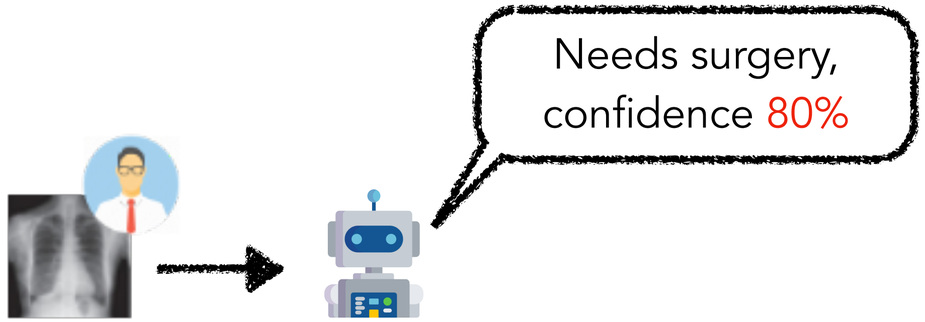
Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.



The doctor decides *optimally*

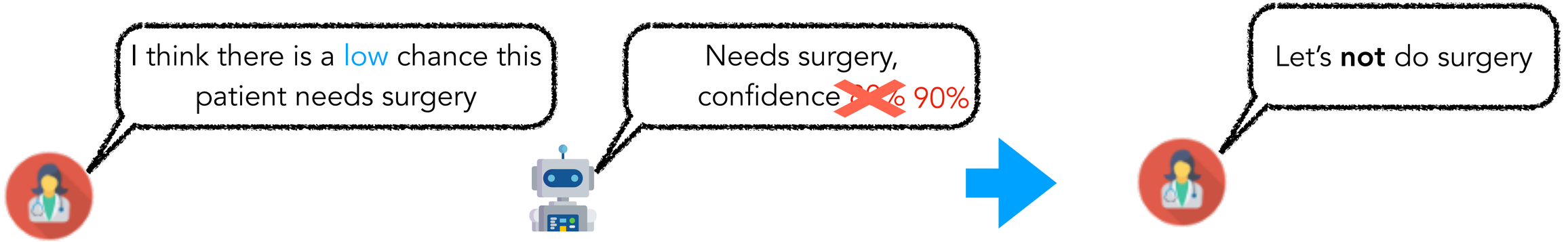
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



Calibration:
Across all patients who  predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

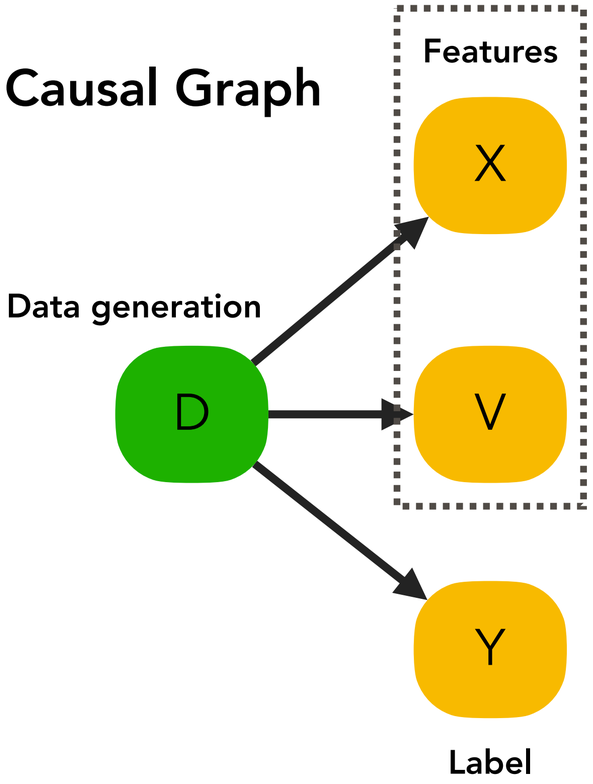
Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic



The doctor decides *optimally*

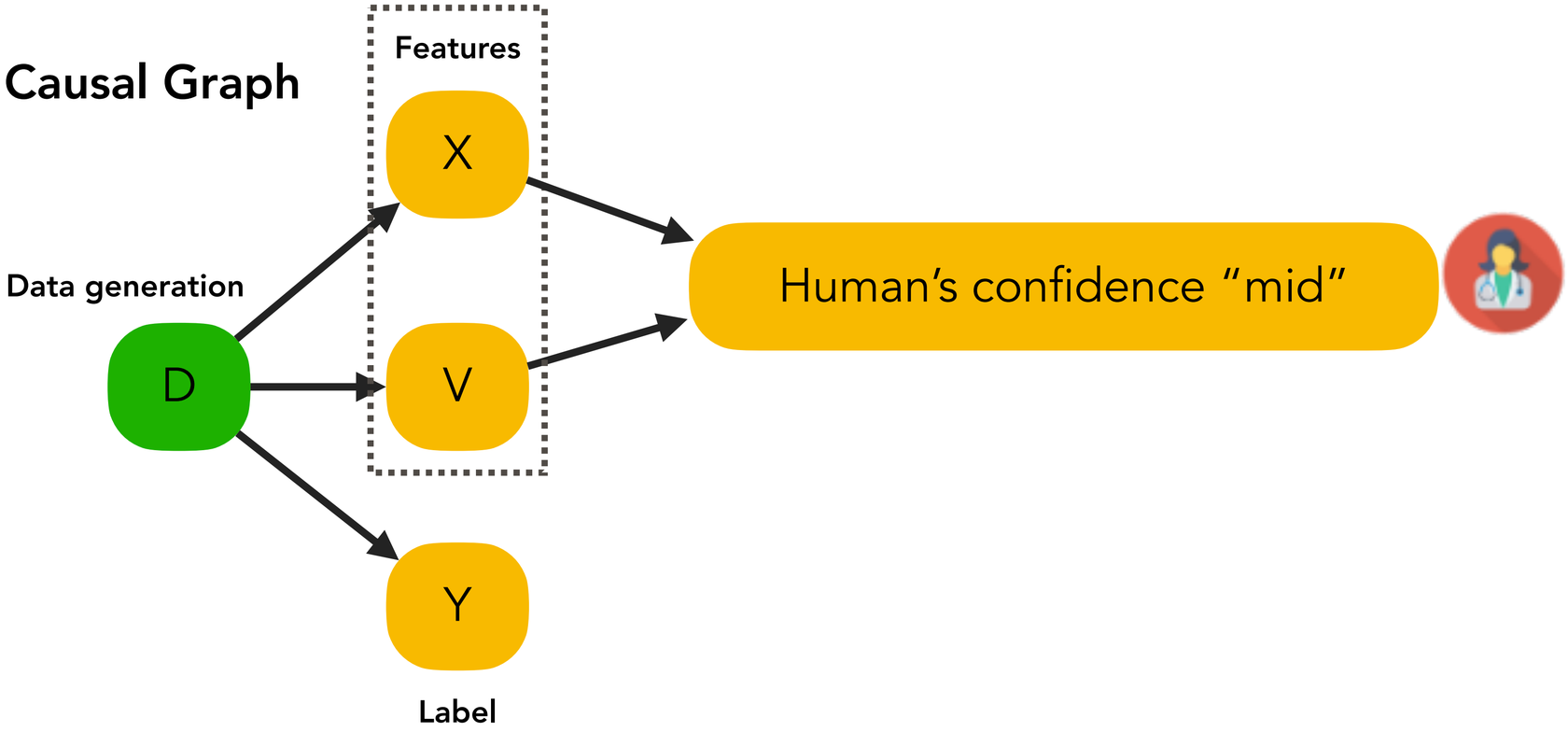
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



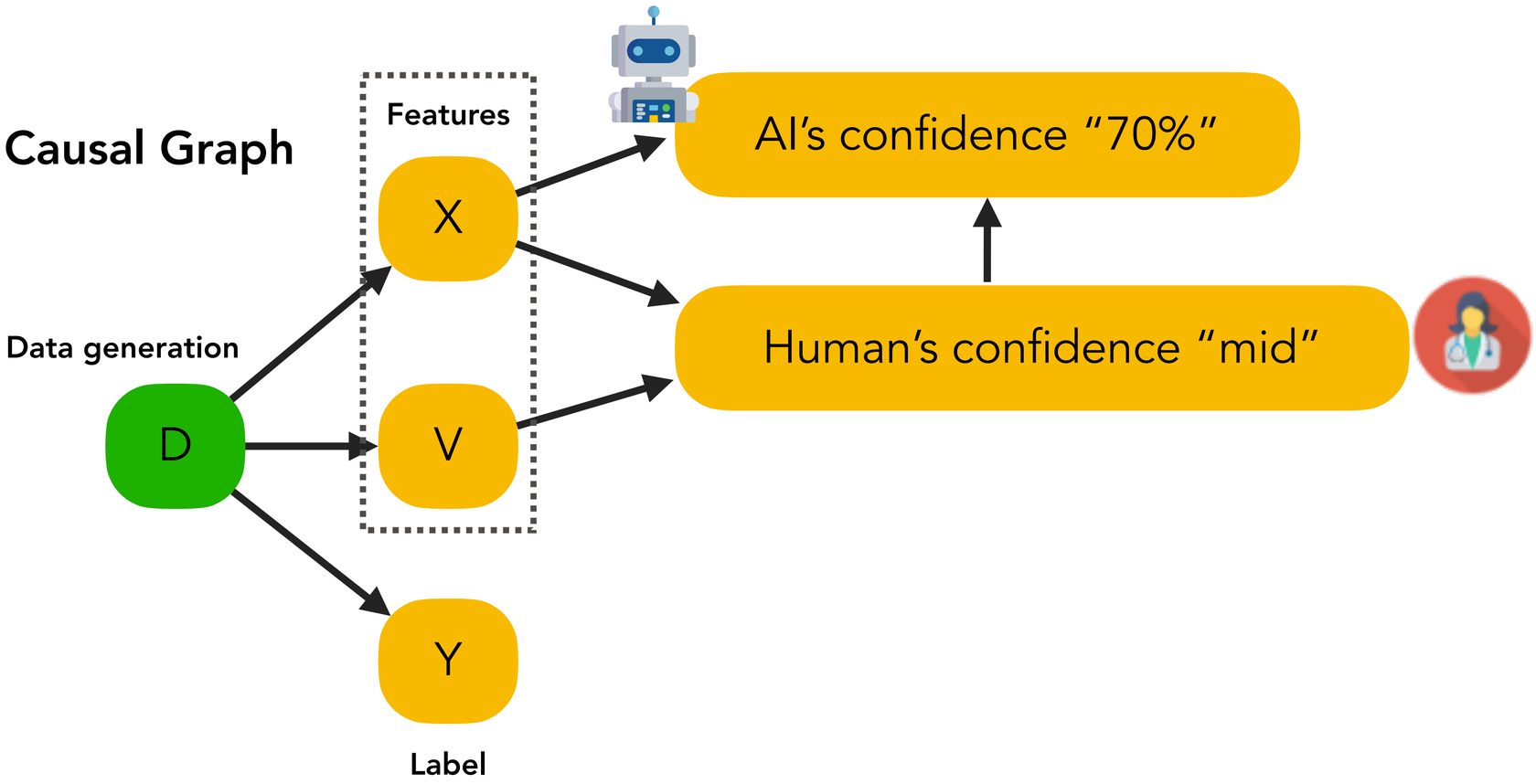
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



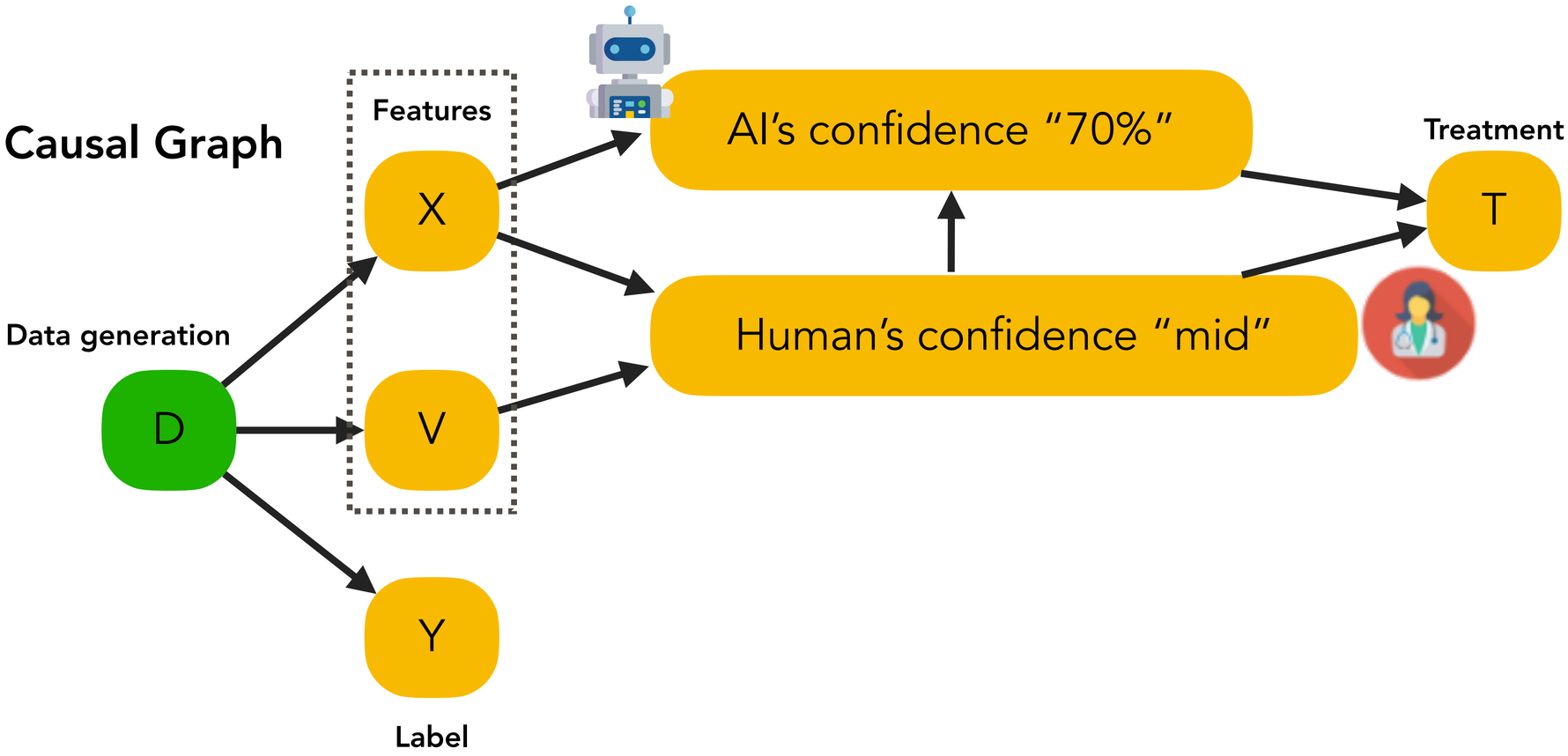
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



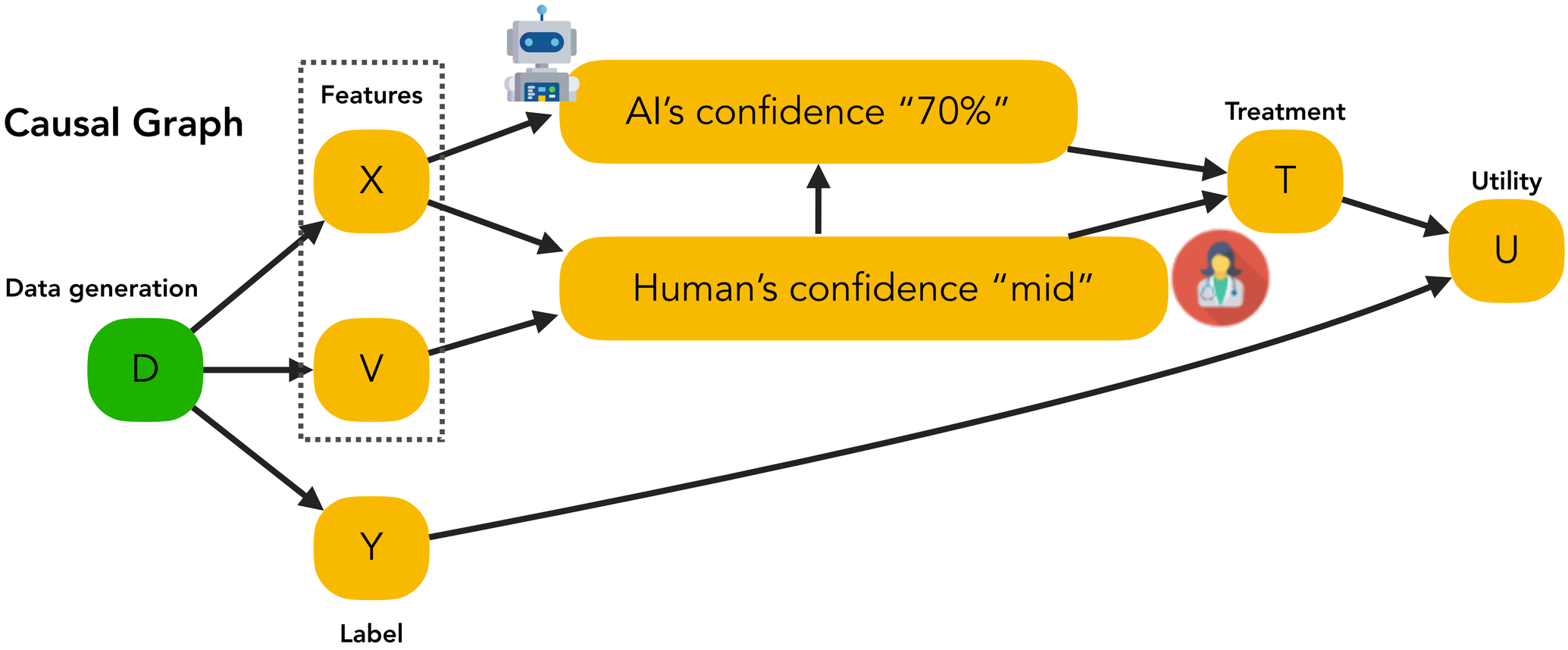
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



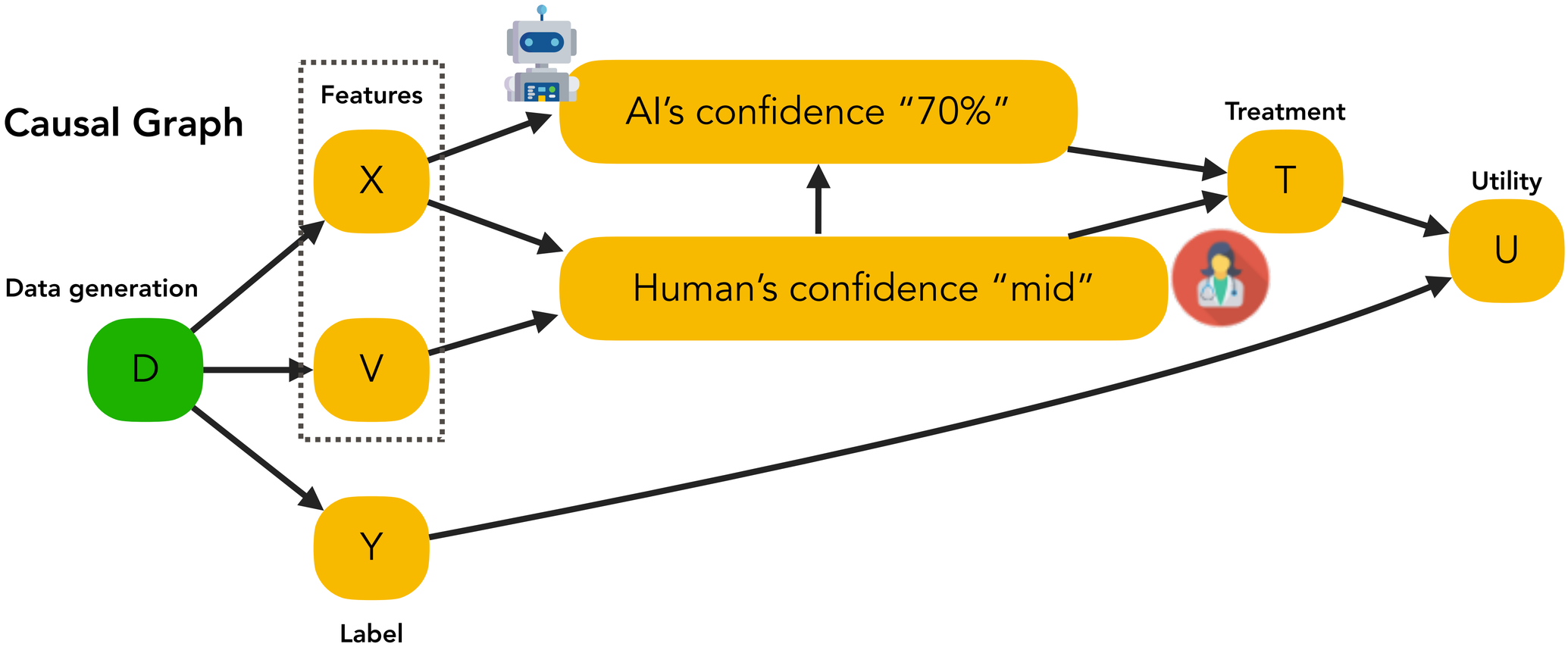
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration

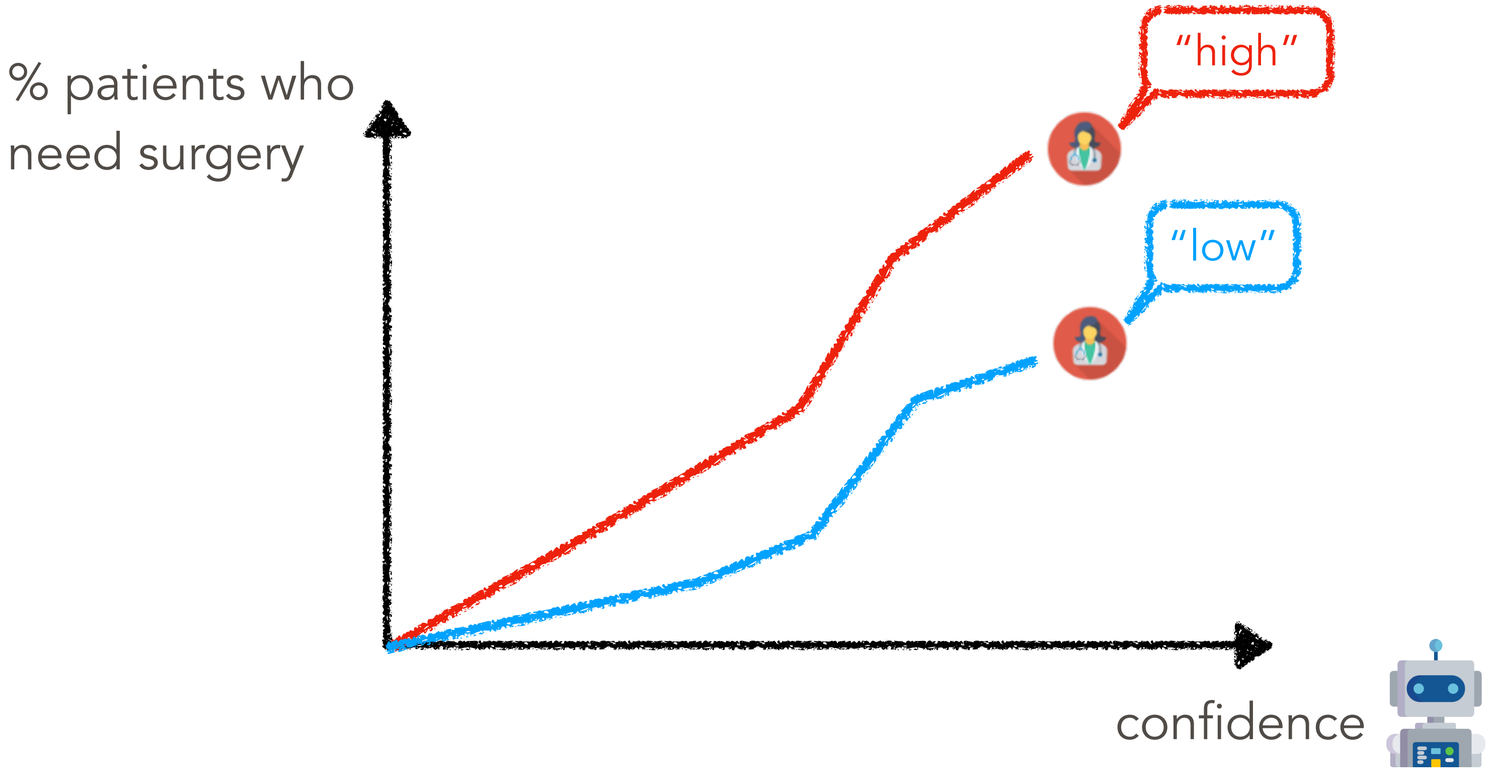


There exist instances of this decision making process in which any monotonic decision policy based on calibrated AI predictions is suboptimal.

Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

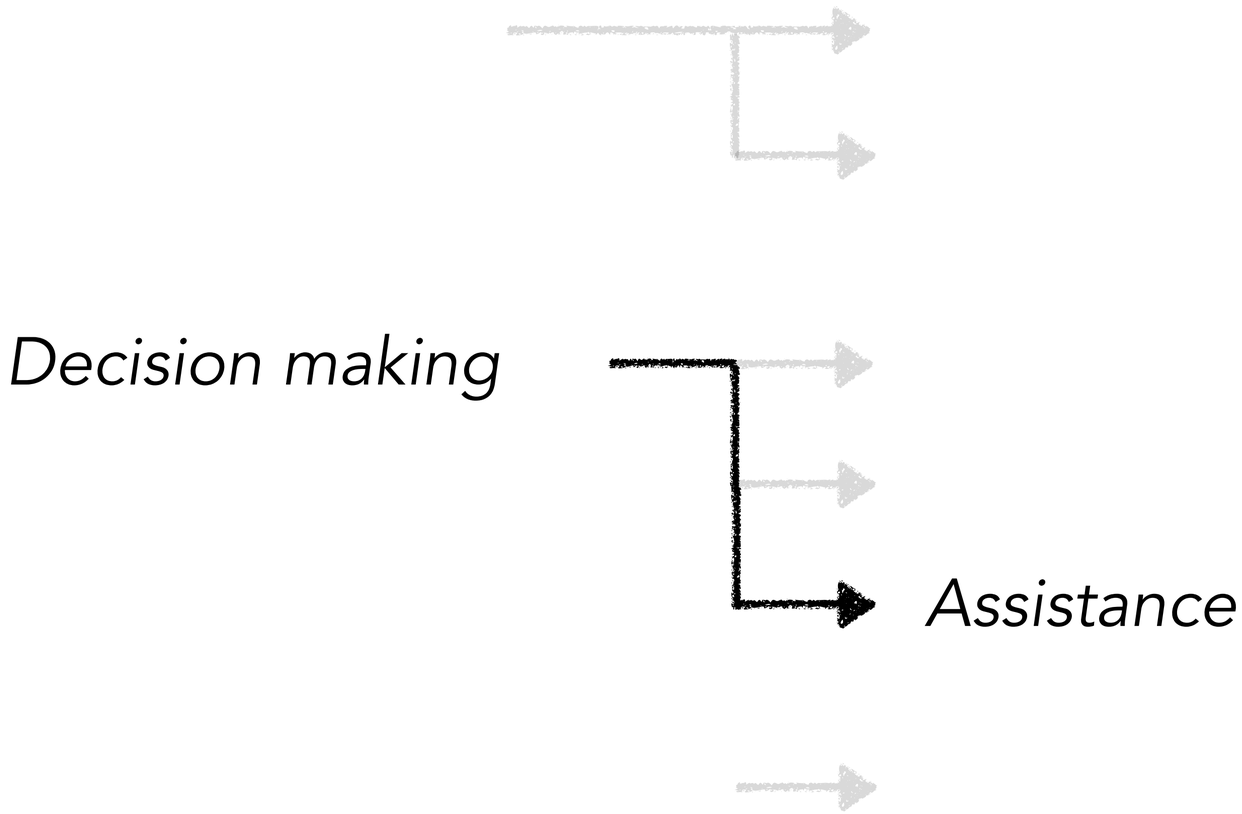
Calibration

To make sure the level of trust the optimal decision maker needs to place on predictions is (always) monotone on the confidence values, one can use **multicalibration.**

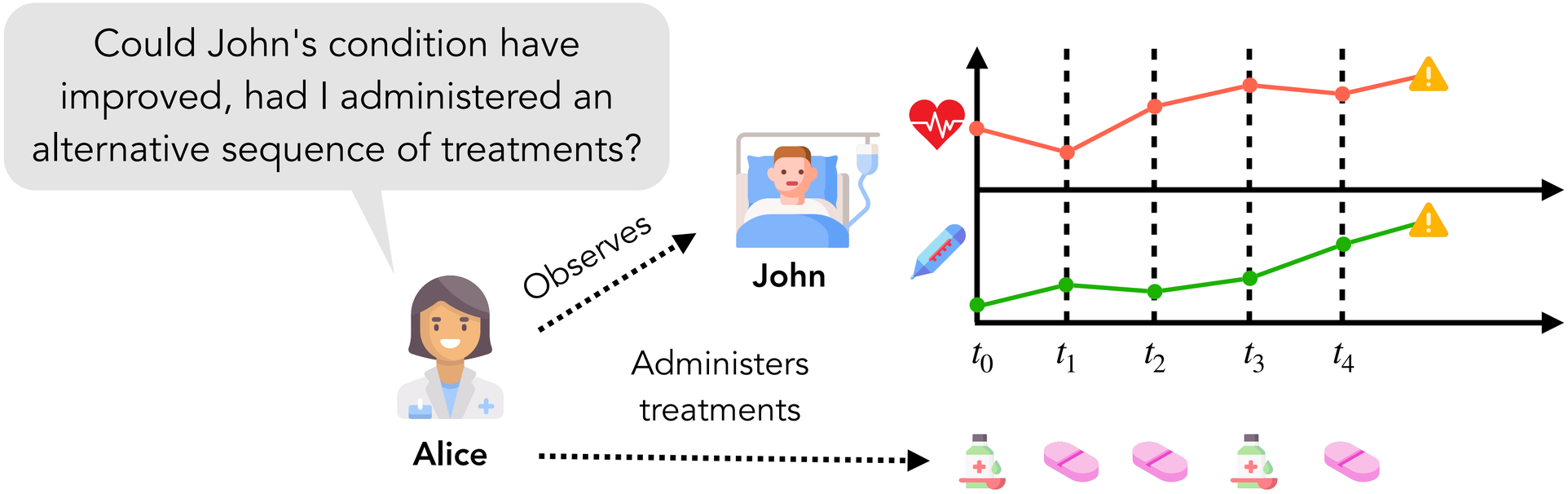


Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Use cases of counterfactuals in machine learning



AI-assisted counterfactuals in sequential decision making



Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

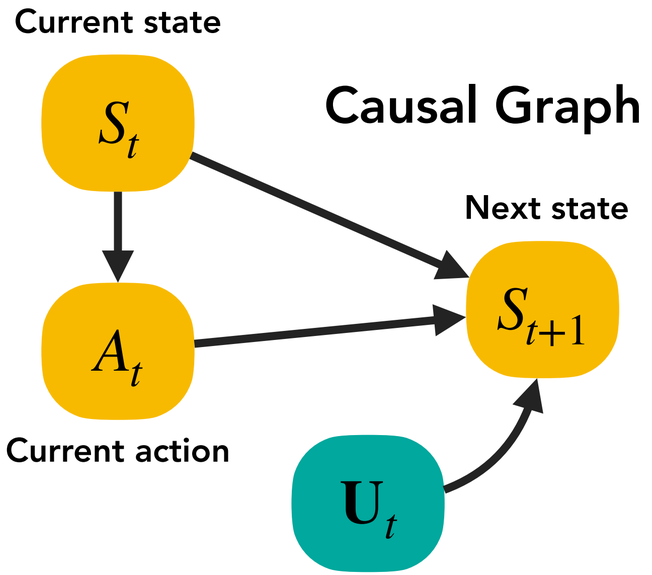
Structural Causal Model $\mathscr{M}$

$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := g_A(S_t, \mathbf{V}_t)$$

$$\mathbf{U}_t \sim P(\mathbf{U})$$

$$\mathbf{V}_t \sim P(\mathbf{V})$$



...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

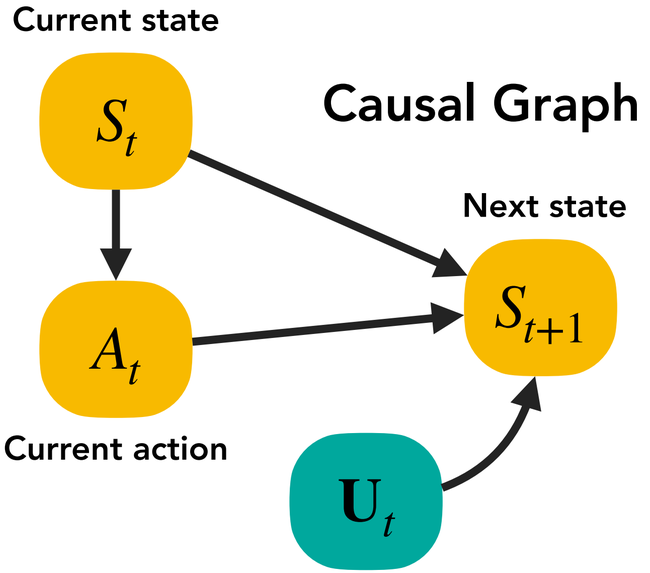
Structural Causal Model $\mathscr{M}$

$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := g_A(S_t, \mathbf{V}_t)$$

$$\mathbf{U}_t \sim P(\mathbf{U})$$

$$\mathbf{V}_t \sim P(\mathbf{V})$$



At state $S_t = s_t$, the doctor took action $A_t = a_t$, what would have happened had the doctor taken action $a' \neq a_t$?

...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

Modified Structural Causal Model $\mathscr{M}_{\{S_t=s_t,\, A_t=a_t\}}$

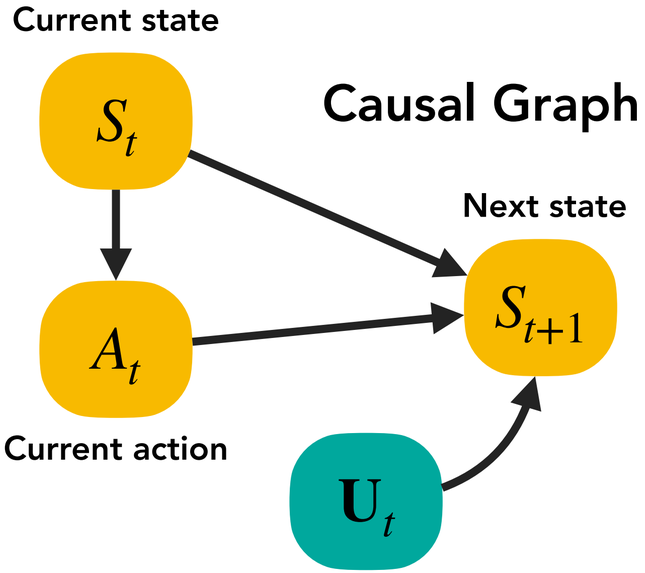
$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := g_A(S_t, \mathbf{V}_t)$$

$$\mathbf{U}_t \sim P(\mathbf{U} \mid S_t = s_t, A_t = a_t)$$

$$\mathbf{V}_t \sim P(\mathbf{V} \mid S_t = s_t)$$

Posterior distribution of the noises



At state $S_t = s_t$, the doctor took action $A_t = a_t$, what would have happened had the doctor taken action $a' \neq a_t$?

...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

Modified Structural Causal Model $\mathscr{M}_{\{S_t=s_t,\, A_t=a_t\}}$

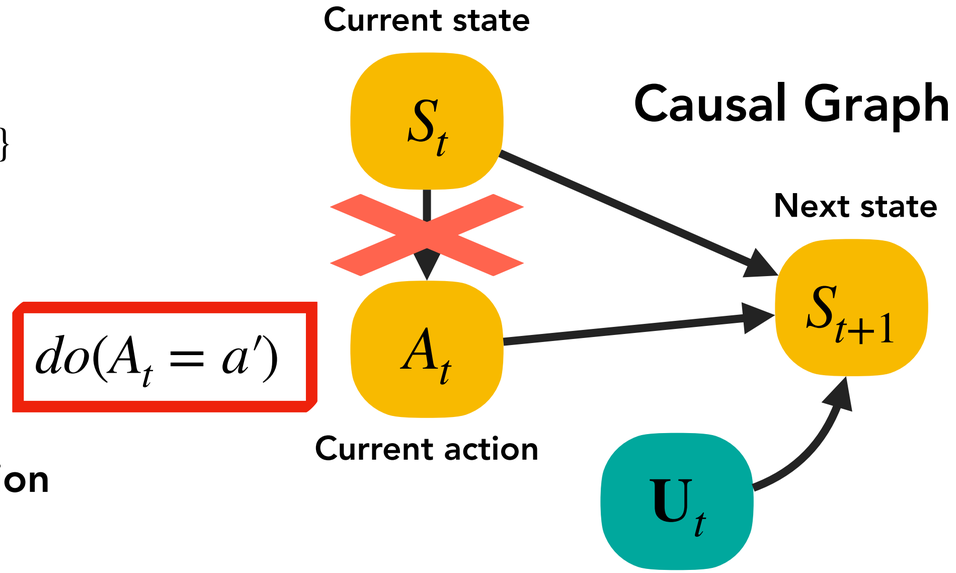
$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := g_A(S_t, \mathbf{V}_t) \qquad A_t := a'$$

$$\mathbf{U}_t \sim P(\mathbf{U} \mid S_t = s_t, A_t = a_t)$$

$$\mathbf{V}_t \sim P(\mathbf{V} \mid S_t = s_t)$$

Posterior distribution of the noises



At state $S_t = s_t$, the doctor took action $A_t = a_t$, what would have happened had the doctor taken action $a' \neq a_t$?

...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

Modified Structural Causal Model $\mathscr{M}_{\{S_t=s_t,\, A_t=a_t\}}$

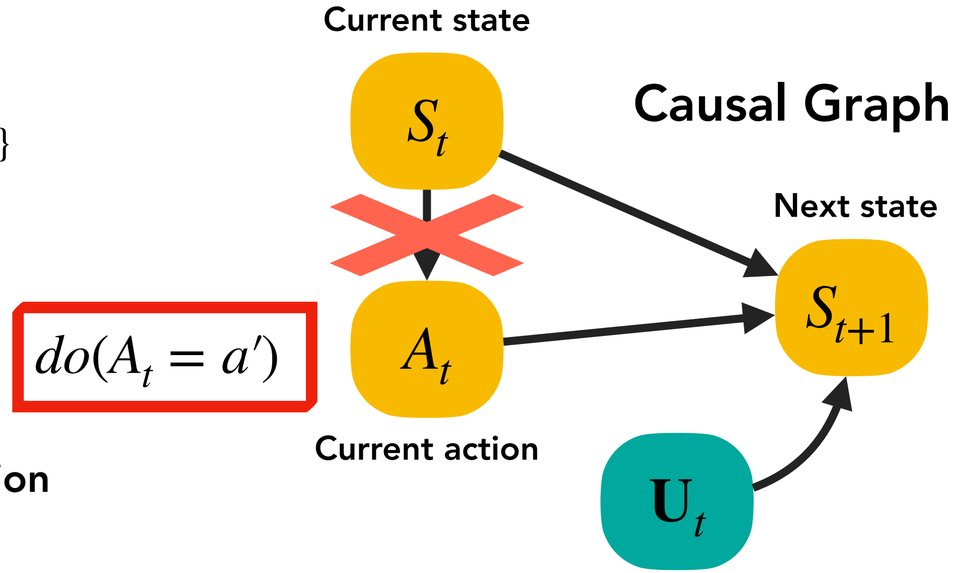
$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := g_A(S_t, \mathbf{V}_t) \qquad A_t := a'$$

$$\mathbf{U}_t \sim P(\mathbf{U} \mid S_t = s_t, A_t = a_t)$$

$$\mathbf{V}_t \sim P(\mathbf{V} \mid S_t = s_t)$$

Posterior distribution of the noises



At state $S_t = s_t$, the doctor took action $A_t = a_t$, what would have happened had the doctor taken action $a' \neq a_t$?

$$S_{t+1} \sim P^{\mathscr{M} \mid S_t=s_t, A_t=a_t\,;\, do(A_t=a')}\left(S_{t+1}\right)$$

...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

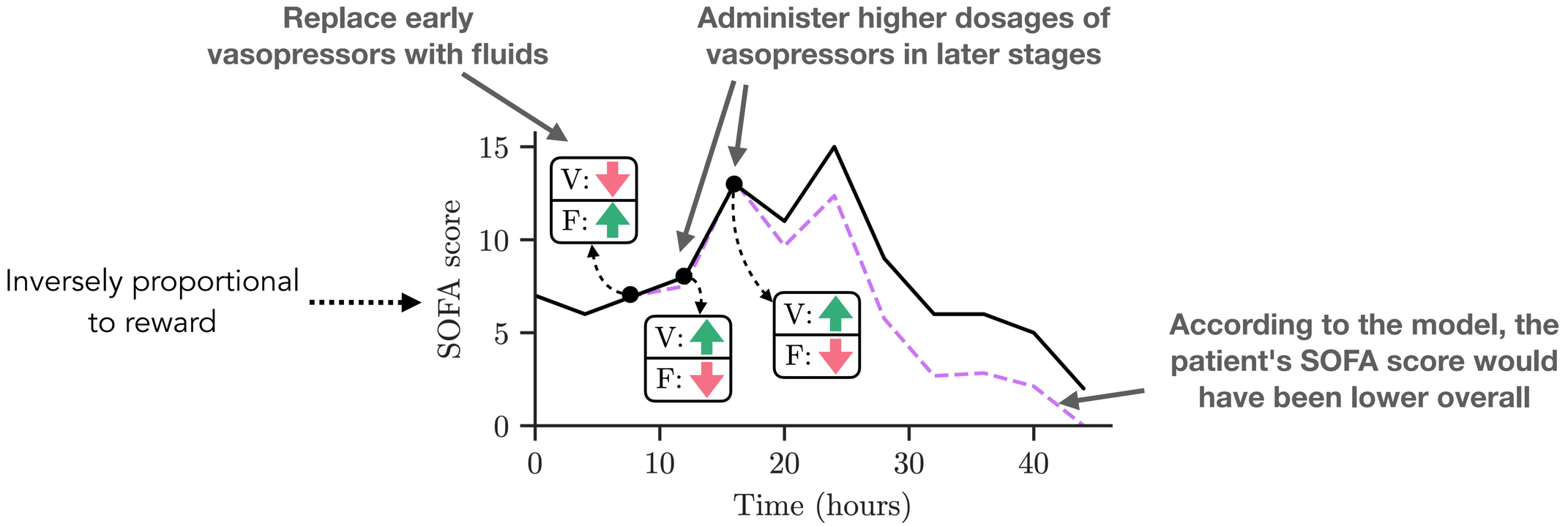
Counterfactually optimal action sequences

Given the counterfactual transition probabilities $S_{t+1} \sim P^{\mathscr{M} \mid S_t=s_t, A_t=a_t\,;\, do(A_t=a')}\left(S_{t+1}\right)$ and a reward function $r(s, a)$, one may find alternative sequence of actions $a'_1, \ldots, a'_{T-1}$ close to the observed actions $a_1, \ldots, a_{T-1}$ that maximizes the average counterfactual reward.

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Counterfactually optimal action sequences

Given the counterfactual transition probabilities $S_{t+1} \sim P^{\mathscr{M} \mid S_t=s_t, A_t=a_t\,;\, do(A_t=a')}\left(S_{t+1}\right)$ and a reward function $r(s, a)$, one may find alternative sequence of actions $a'_1, \ldots, a'_{T-1}$ close to the observed actions $a_1, \ldots, a_{T-1}$ that maximizes the average counterfactual reward.



Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Use cases of counterfactuals in machine learning

Reinforcement learning → *Training*

Counterfactually-guided training in reinforcement learning

In reinforcement learning, given a transition probability $P(s' \mid s, a)$ and a reward function $r(s, a)$, the goal is to design an action policy $a := \pi(s)$ with the highest average reward, i.e.

$$\pi^* = \text{argmax}_\pi \mathbb{E}_{\tau \sim \pi, P} \left[R(\tau) \right] \quad \text{where} \quad R(\tau) = \sum_{t=1}^{T} R(s_t, a_t)$$

Buesing et al. "*Woulda, coulda, shoulda: Counterfactually-guided policy search.*" ICLR, 2018.

Counterfactually-guided training in reinforcement learning

In reinforcement learning, given a transition probability $P(s' \mid s, a)$ and a reward function $r(s, a)$, the goal is to design an action policy $a := \pi(s)$ with the highest average reward, i.e.

$$\pi^* = \text{argmax}_\pi \mathbb{E}_{\tau \sim \pi, P} \left[R(\tau) \right] \quad \text{where} \quad R(\tau) = \sum_{t=1}^{T} R(s_t, a_t)$$

Counterfactually-guided training refers to the evaluation of the above expectation using data gathered via an action policy $\pi' \neq \pi$ and counterfactual reasoning

Buesing et al. "*Woulda, coulda, shoulda: Counterfactually-guided policy search.*" ICLR, 2018.

Counterfactually-guided training in reinforcement learning

In reinforcement learning, given a transition probability $P(s' \mid s, a)$ and a reward function $r(s, a)$, the goal is to design an action policy $a := \pi(s)$ with the highest average reward, i.e.

$$\pi^* = \text{argmax}_\pi \mathbb{E}_{\tau \sim \pi, P}\left[R(\tau)\right] \quad \text{where} \quad R(\tau) = \sum_{t=1}^{T} R(s_t, a_t)$$

Counterfactually-guided training refers to the evaluation of the above expectation using data gathered via an action policy $\pi' \neq \pi$ and counterfactual reasoning

Structural Causal Model $\mathscr{M}$

$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := \pi'(S_t)$$

$$\mathbf{U}_t \sim P(\mathbf{U})$$

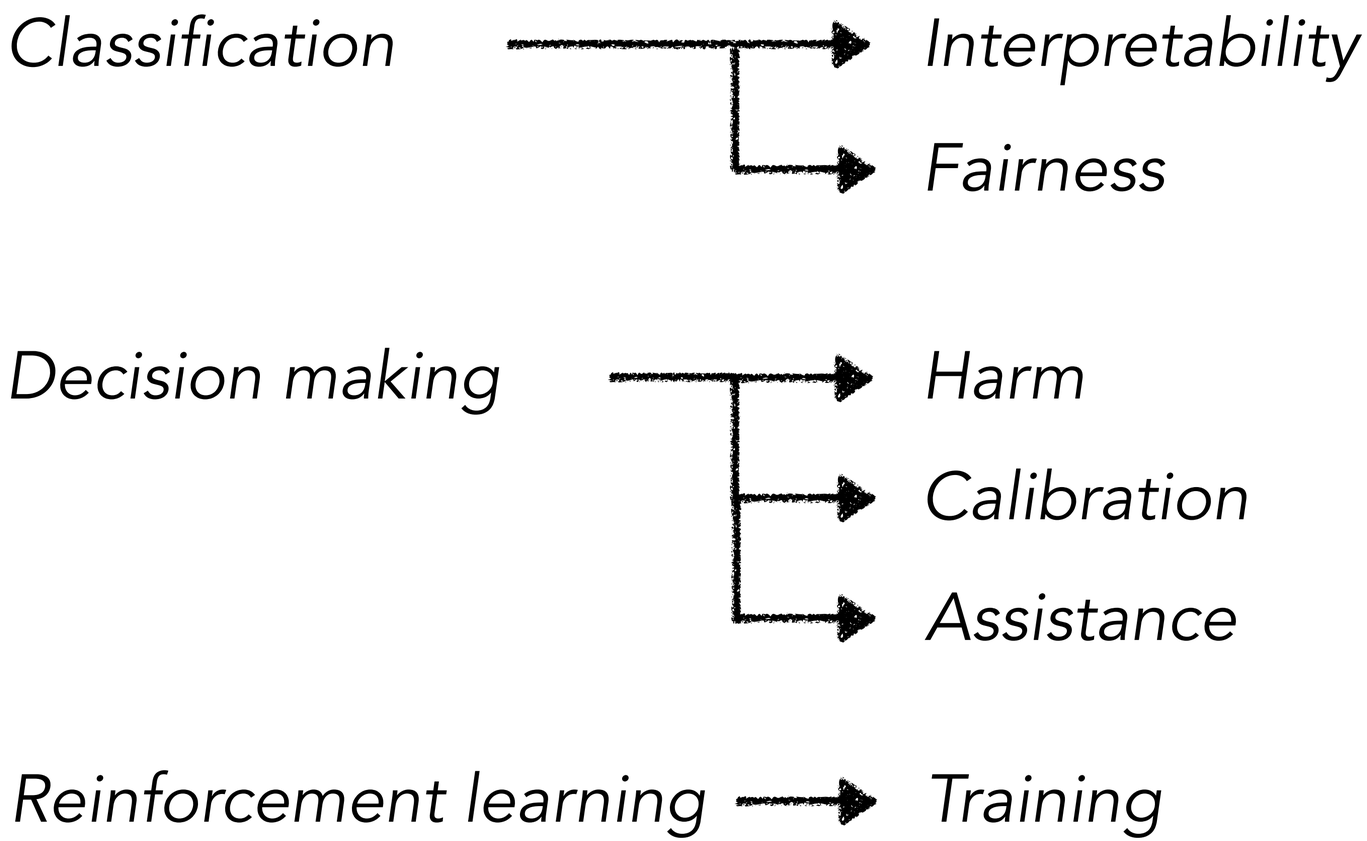
Key idea:

$$E_{S_t,\, a_t \sim P^{\mathscr{M}}}\left[P^{\mathscr{M} \mid S_t=s_t, A_t=a_t\,;\, do(A_t=\pi(S_t))}\right] = P^{\mathscr{M}\,;\, do(A_t=\pi(S_t))}$$

Observational probability | **Counterfactual probability** | **Interventional probability**

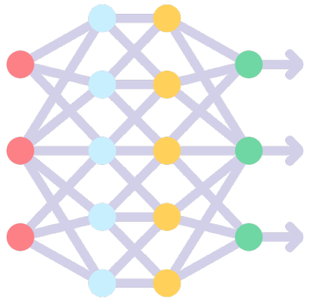
Buesing et al. "*Woulda, coulda, shoulda: Counterfactually-guided policy search.*" ICLR, 2018.

Use cases of counterfactuals in machine learning



Outline

Cognitive science



Machine learning



Large language models

Outline

Cognitive science

Machine learning



Large language models

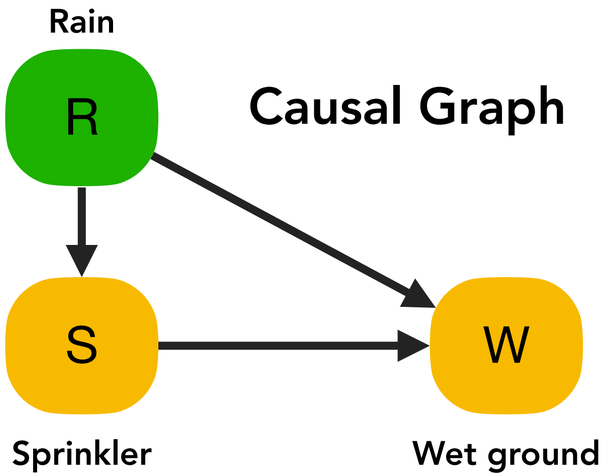
Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.

Jin et al. "*Cladder: Assessing causal reasoning in language models.*" NeurIPS, 2023.
Kiciman et al. "*Causal reasoning and large language models: Opening a new frontier for causality.*" TMLR, 2023.
Zhang et al. "*What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models.*" CVPR, 2024.

Can LLMs reason about counterfactuals?

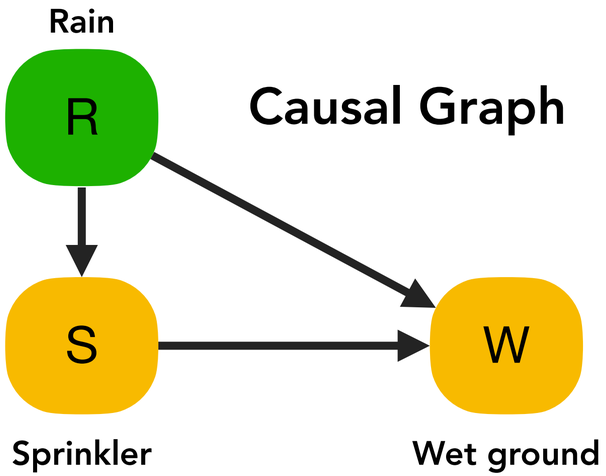
Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.



Jin et al. "*Cladder: Assessing causal reasoning in language models.*" NeurIPS, 2023.
Kiciman et al. "*Causal reasoning and large language models: Opening a new frontier for causality.*" TMLR, 2023.
Zhang et al. "*What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models.*" CVPR, 2024.

Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.



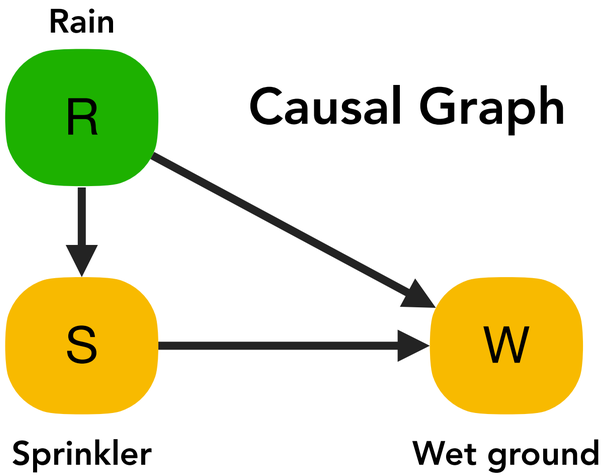
Verbal query

We know that rainy season causes rain and sprinkler off. Rain or sprinkler on causes wet ground. Would the ground be wet if it was the dry season instead of the rainy season?

Jin et al. "*Cladder: Assessing causal reasoning in language models.*" NeurIPS, 2023.
Kiciman et al. "*Causal reasoning and large language models: Opening a new frontier for causality.*" TMLR, 2023.
Zhang et al. "*What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models.*" CVPR, 2024.

Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.



Causal Graph

Verbal query

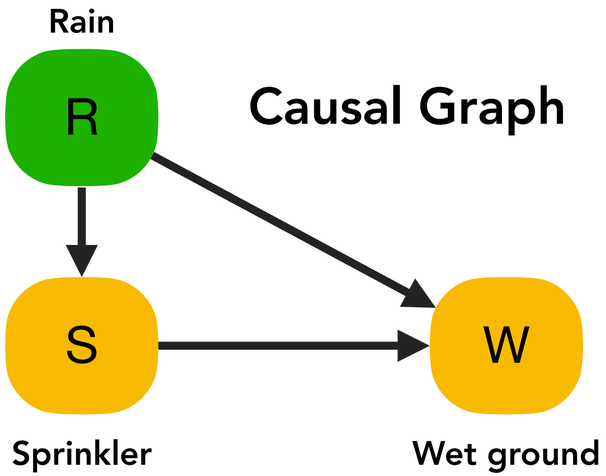
We know that rainy season causes rain and sprinkler off. Rain or sprinkler on causes wet ground. Would the ground be wet if it was the dry season instead of the rainy season?

Comparison of the LLM's response with ground truth

Jin et al. "*Cladder: Assessing causal reasoning in language models.*" NeurIPS, 2023.
Kiciman et al. "*Causal reasoning and large language models: Opening a new frontier for causality.*" TMLR, 2023.
Zhang et al. "*What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models.*" CVPR, 2024.

Can LLMs reason about counterfactuals?

Several studies have looked at the ability of large language models to provide responses to interventional and counterfactual questions.



Verbal query

We know that rainy season causes rain and sprinkler off. Rain or sprinkler on causes wet ground. Would the ground be wet if it was the dry season instead of the rainy season?

Comparison of the LLM's response with ground truth

	Overall Acc.	Acc. by Rung		
		1	2	3
Random	49.27	50.28	48.40	49.12
LLaMa	44.03	48.23	29.46	52.66
Alpaca	44.66	52.03	29.53	51.13
GPT-3 Non-Instr. (davinci)	49.92	50.00	49.75	50.00
GPT-3 Instr. (text-davinci-001)	51.40	51.30	52.63	50.47
GPT-3 Instr. (text-davinci-002)	53.15	50.85	56.96	51.90
GPT-3 Instr. (text-davinci-003)	56.26	51.11	62.97	54.96
GPT-3.5	52.18	51.80	54.78	50.32
GPT-4	62.03	63.01	62.82	60.55
+ CausalCoT	**70.40**	**83.35**	**67.47**	**62.05**

Jin et al. "*Cladder: Assessing causal reasoning in language models.*" NeurIPS, 2023.
Kiciman et al. "*Causal reasoning and large language models: Opening a new frontier for causality.*" TMLR, 2023.
Zhang et al. "*What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models.*" CVPR, 2024.

Deep dive:
Can LLMs reason about counterfactuals of their own outputs?

Chatzi et al. "*Counterfactual token generation in large language models.*" CLeaR, 2025.

Autoregressive token generation

Prompt: What is your favorite color?
Response:

Autoregressive token generation

Prompt: What is your favorite color?
Response: My

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the

Autoregressive token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.

Autoregressive token generation

Prompt: What is your favorite color?
Response: I like pink. Do you like it as well?

Prompt: What is your favorite color?
Response: My favorite color is blue. It is the color of the sea.

Autoregressive token generation

Prompt: What is your favorite color?
Response: I like pink. Do you like it as well?

Prompt: What is your favorite color?
Response: My favorite color is blue. It is the color of the sea.

Prompt: What is your favorite color?
Response: I don’t really like colors. Black is fine I guess…

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.

Factual generation

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.



Factual generation

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.



Factual generation

Prompt: What is your favorite color?
Response: My favorite color is **green.**

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.



Factual generation

Prompt: What is your favorite color?
Response: My favorite color is **green.**
It is the color of the trees.

Counterfactual generation

Factual and counterfactual token generation

Prompt: What is your favorite color?
Response: My favorite color is blue.
It is the color of the sea.



Factual generation

Prompt: What is your favorite color?
Response: My favorite color is **green.**
It is the color of the trees.

Counterfactual generation

Factual and counterfactual worlds

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response:

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man working as a software engineer.

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man working as a software engineer.

Factual world

Factual and counterfactual worlds

Prompt: Generate a fictional employee
Response: Jamie is a 28-year old man working as a software engineer.



Factual world

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man working as a software engineer.



Factual world

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman**

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man working as a software engineer.



Factual world

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman** working as a software engineer.

Counterfactual world

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man working as a software engineer.



Factual world

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman** working as a software engineer.

Counterfactual world

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man working as a software engineer.



Factual world

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman** working as a software engineer.

Counterfactual world

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman** working as a nurse.

Another counterfactual world…

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man working as a software engineer.

Factual world



Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman** working as a software engineer.

Counterfactual world

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman** working as a nurse.

Another counterfactual world…

😕

Performing this type of analysis is **not possible** with vanilla implementations of LLMs…

Factual and counterfactual worlds

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old man working as a software engineer.



Factual world

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman** working as a software engineer.

Counterfactual world

Prompt: Generate a fictional employee.
Response: Jamie is a 28-year old **woman** working as a nurse.

Another counterfactual world…

😕

Performing this type of analysis is **not possible** with vanilla implementations of LLMs…

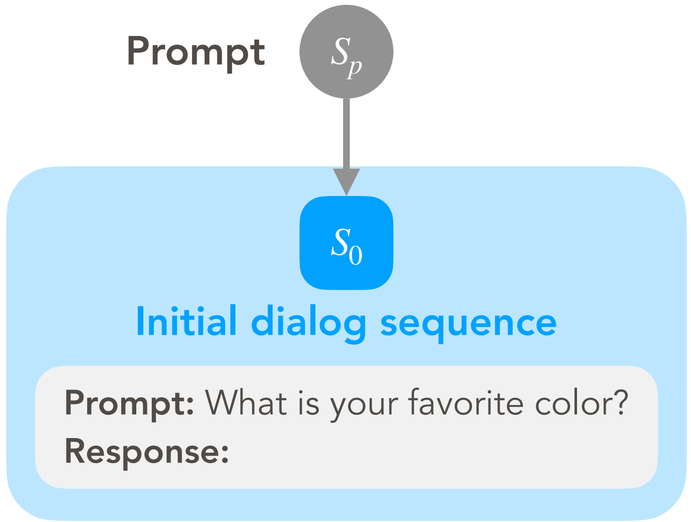
😀

We we will see a method based on **structural causal models** that makes it possible

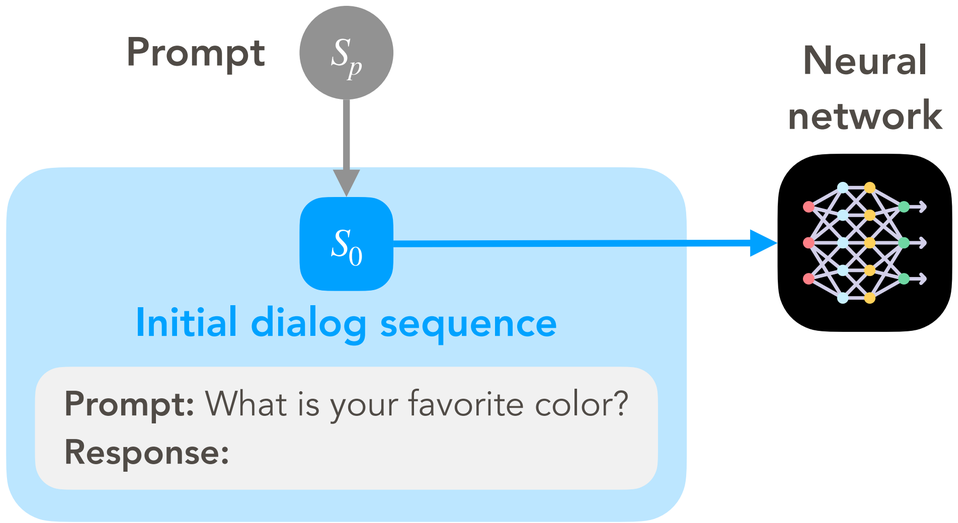
Autoregressive generation through a causal lens

Prompt S_p

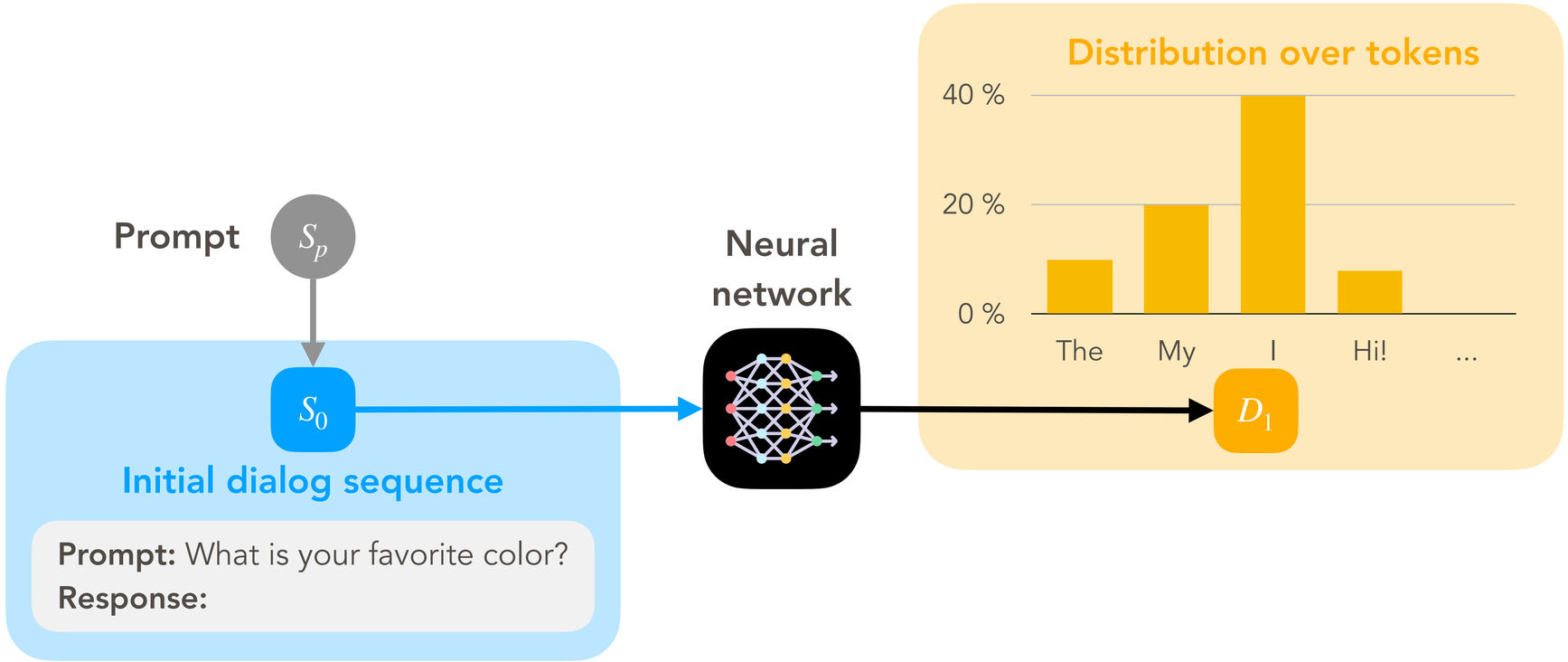
Autoregressive generation through a causal lens



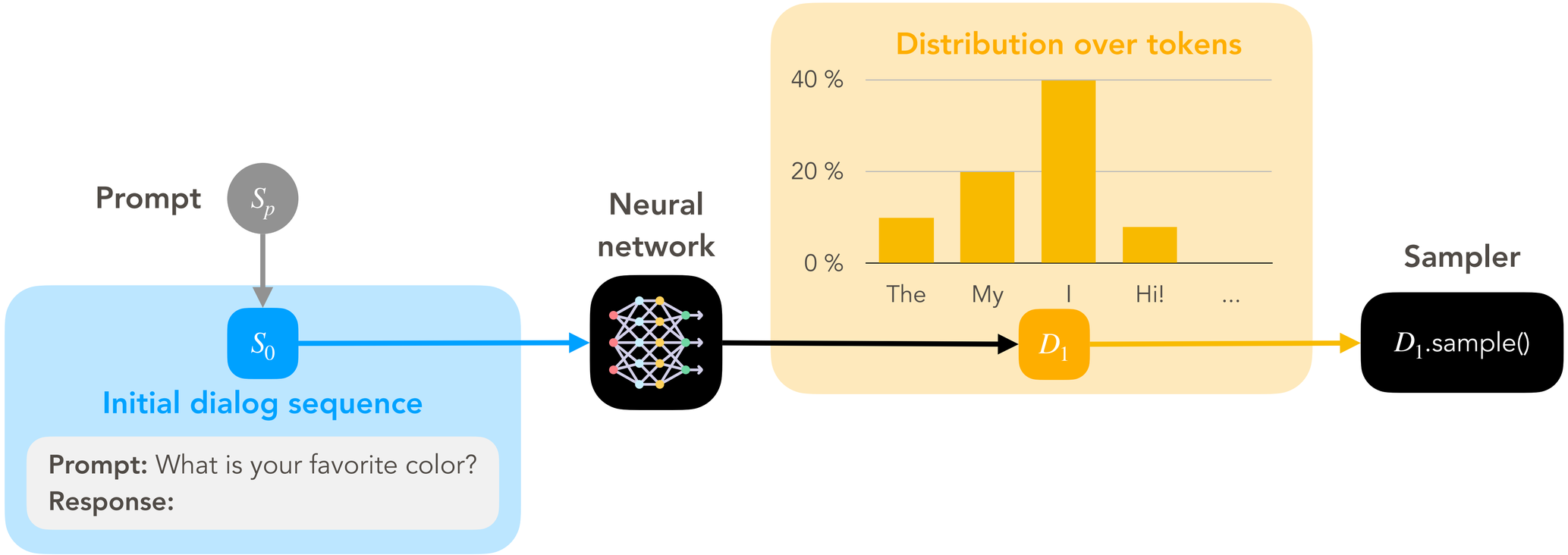
Autoregressive generation through a causal lens



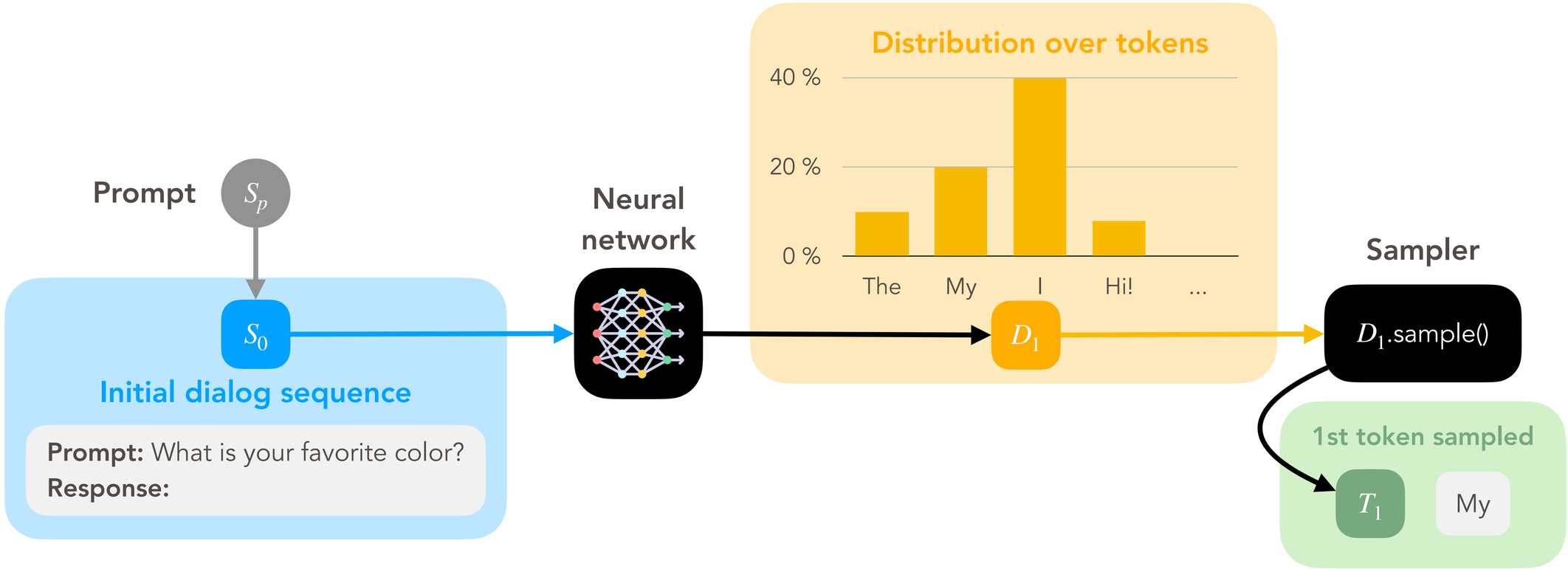
Autoregressive generation through a causal lens



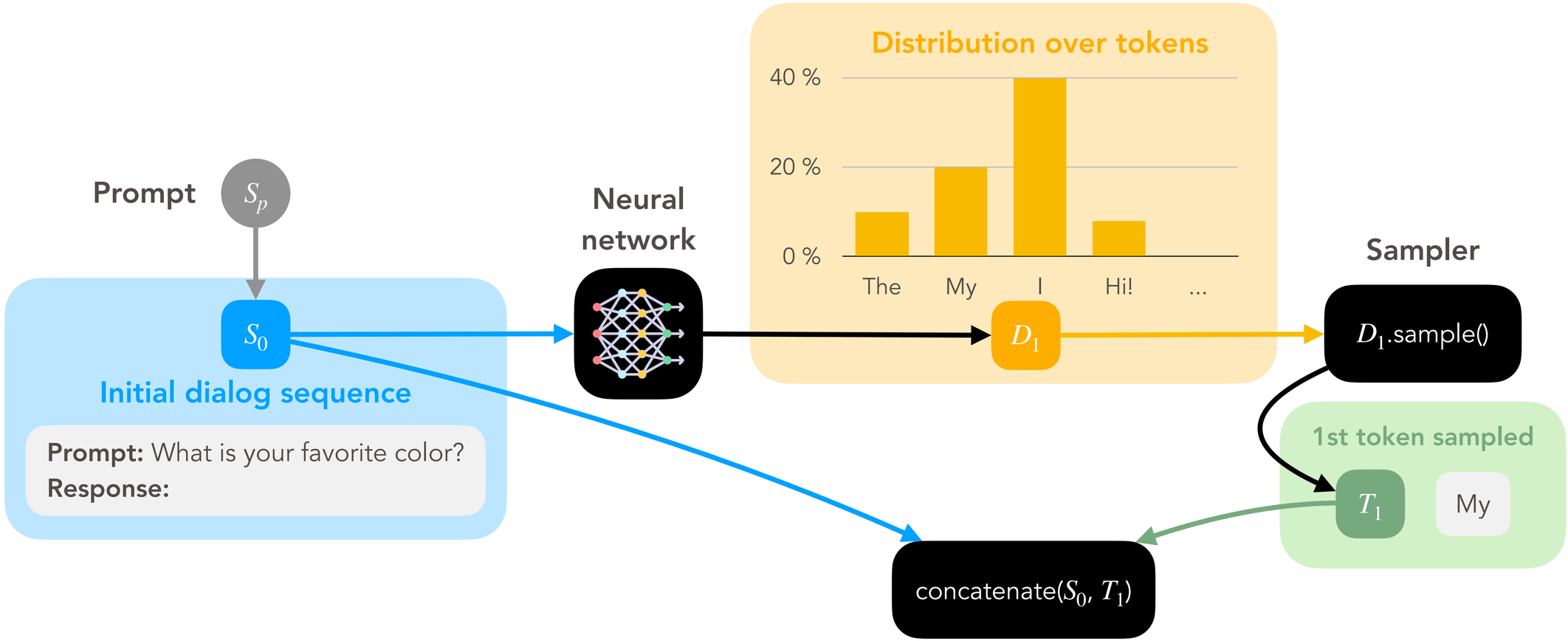
Autoregressive generation through a causal lens



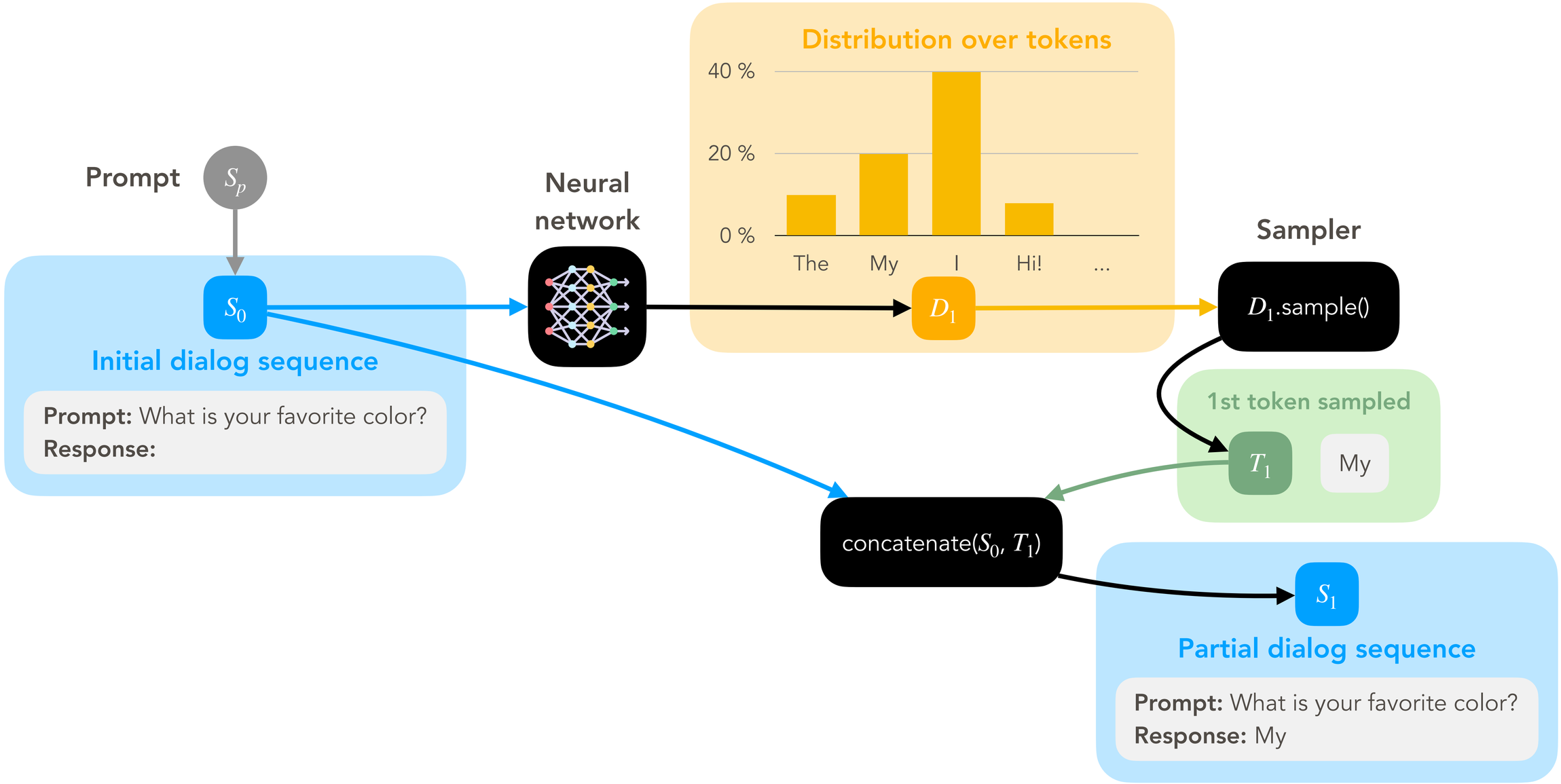
Autoregressive generation through a causal lens



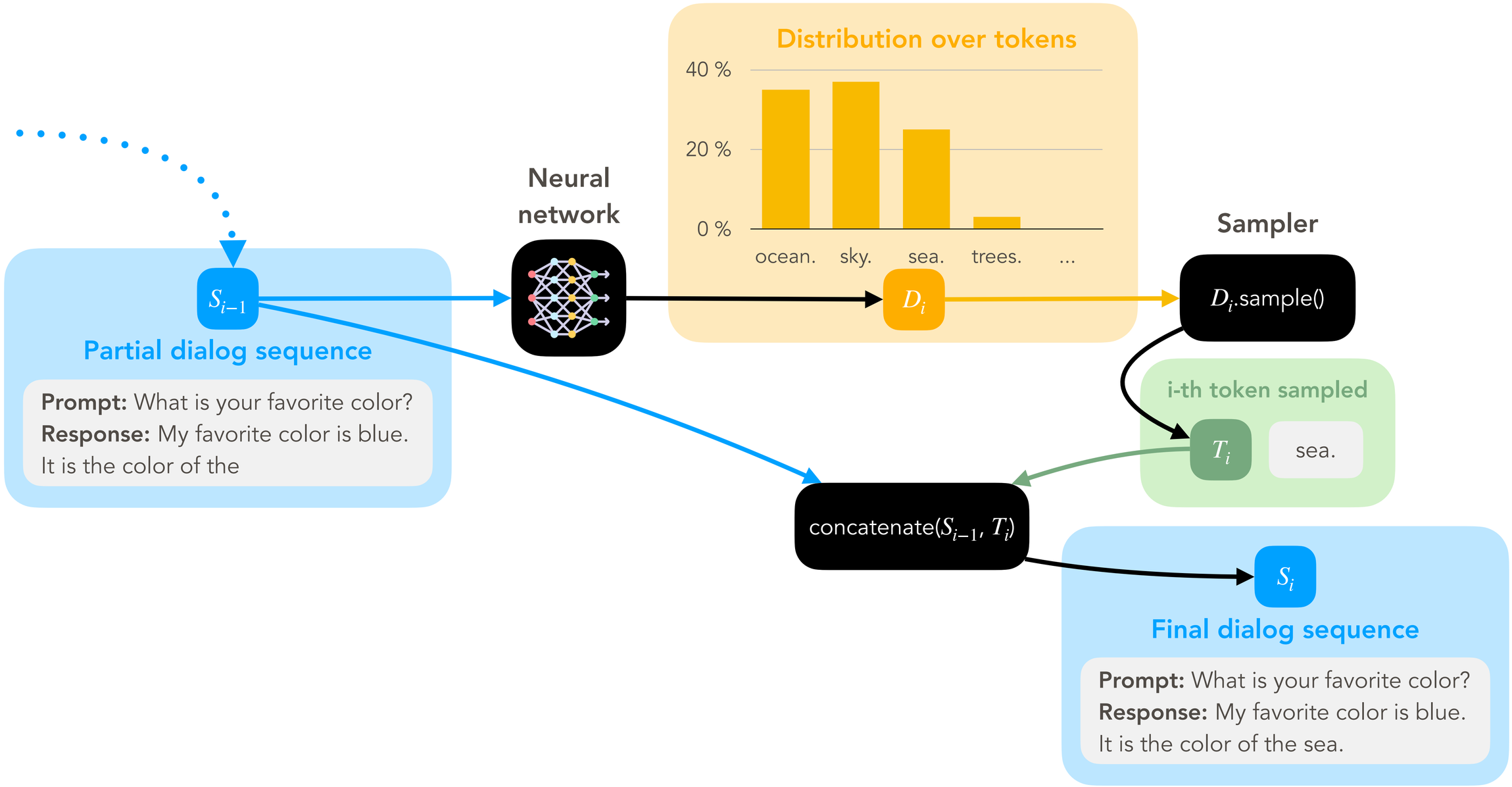
Autoregressive generation through a causal lens



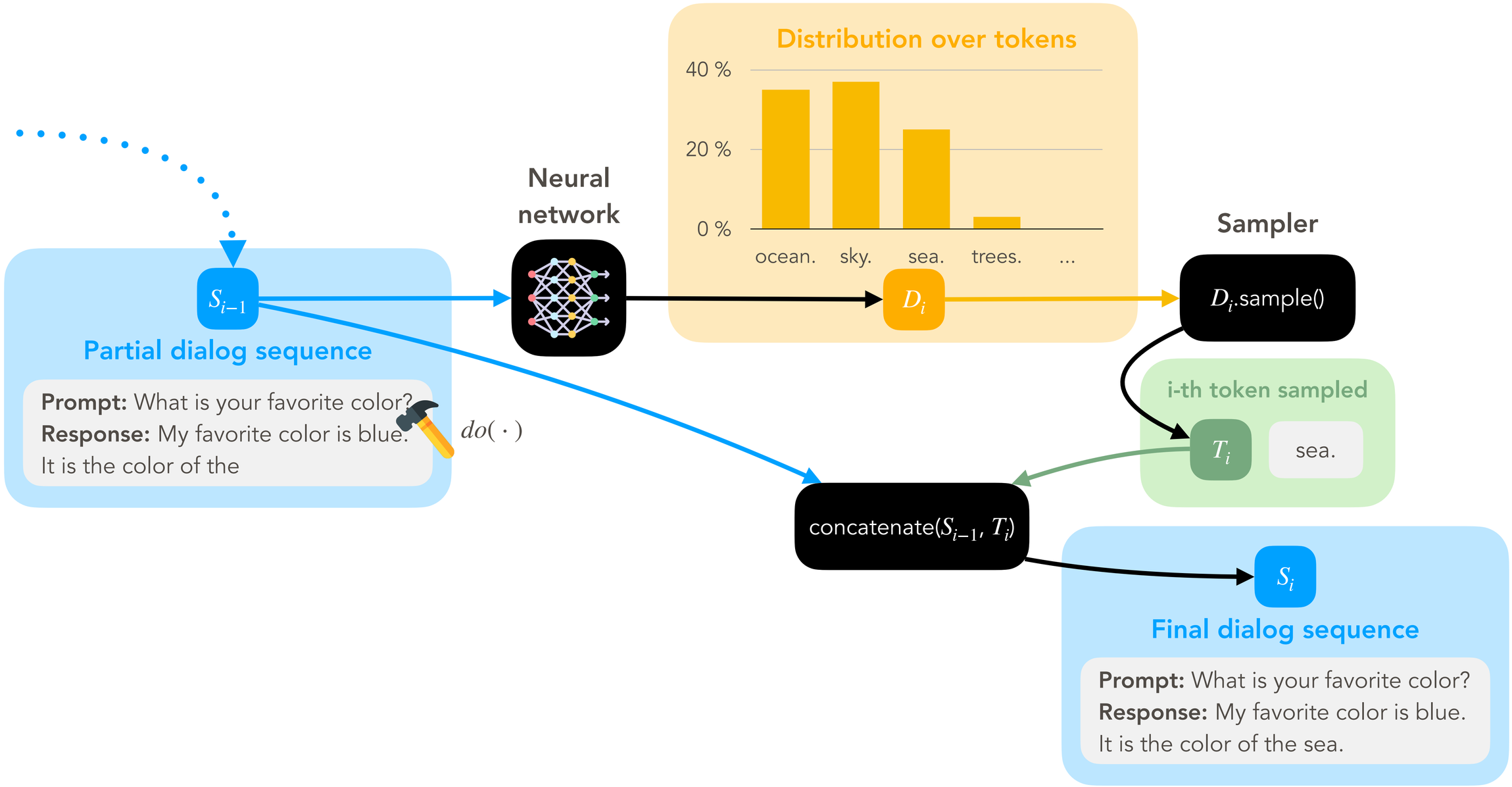
Autoregressive generation through a causal lens



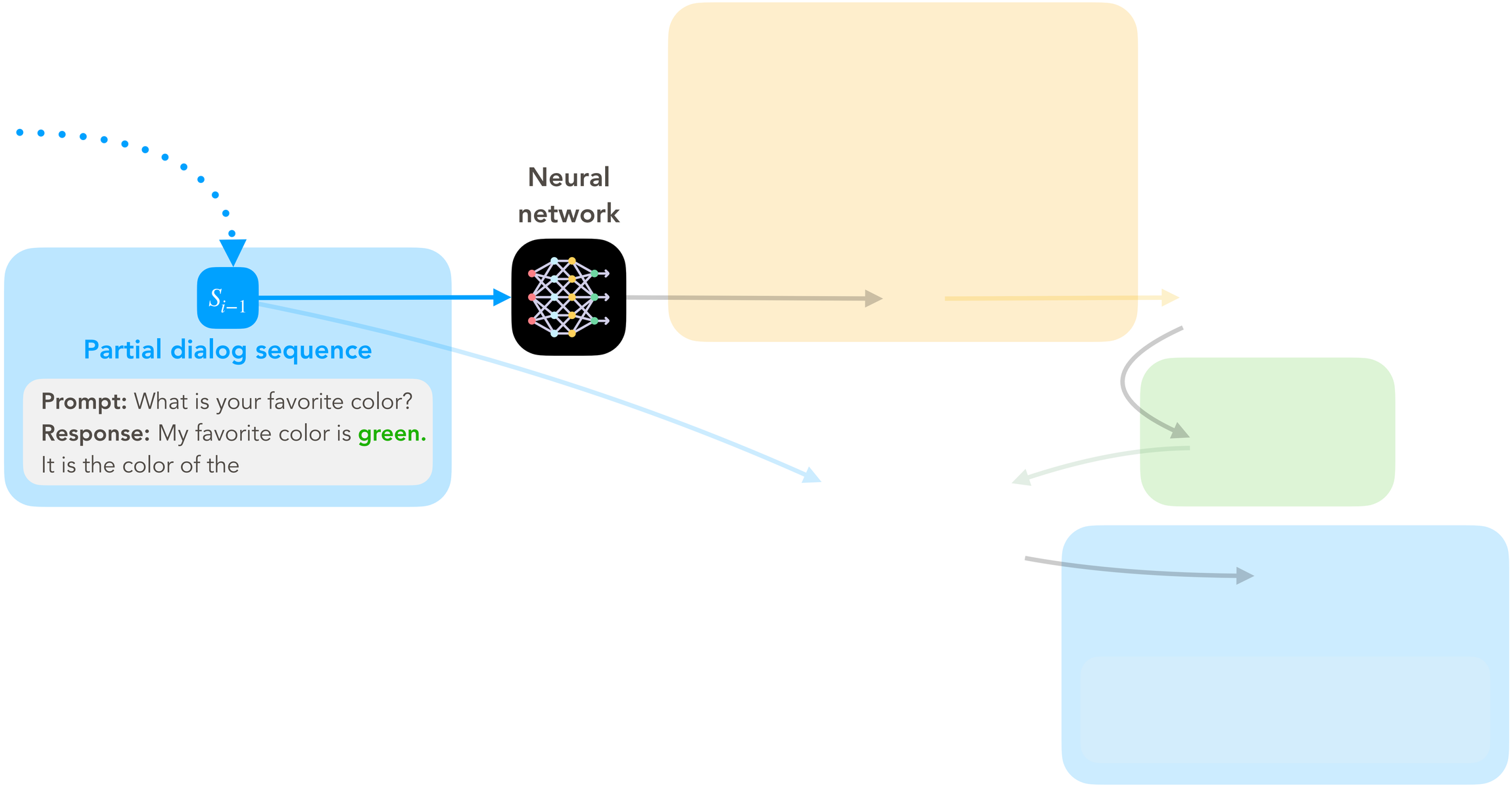
Autoregressive generation through a causal lens



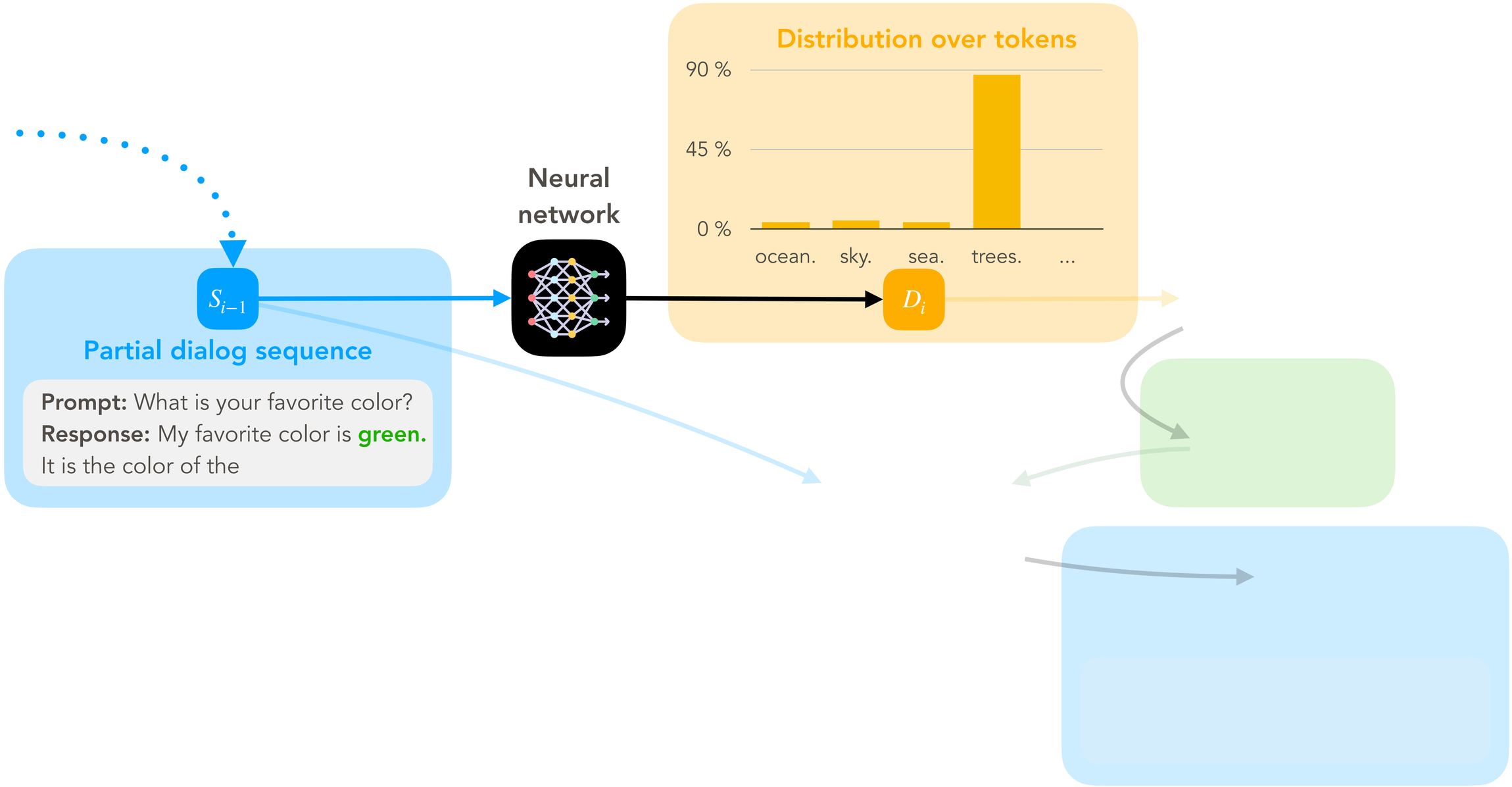
Autoregressive generation through a causal lens



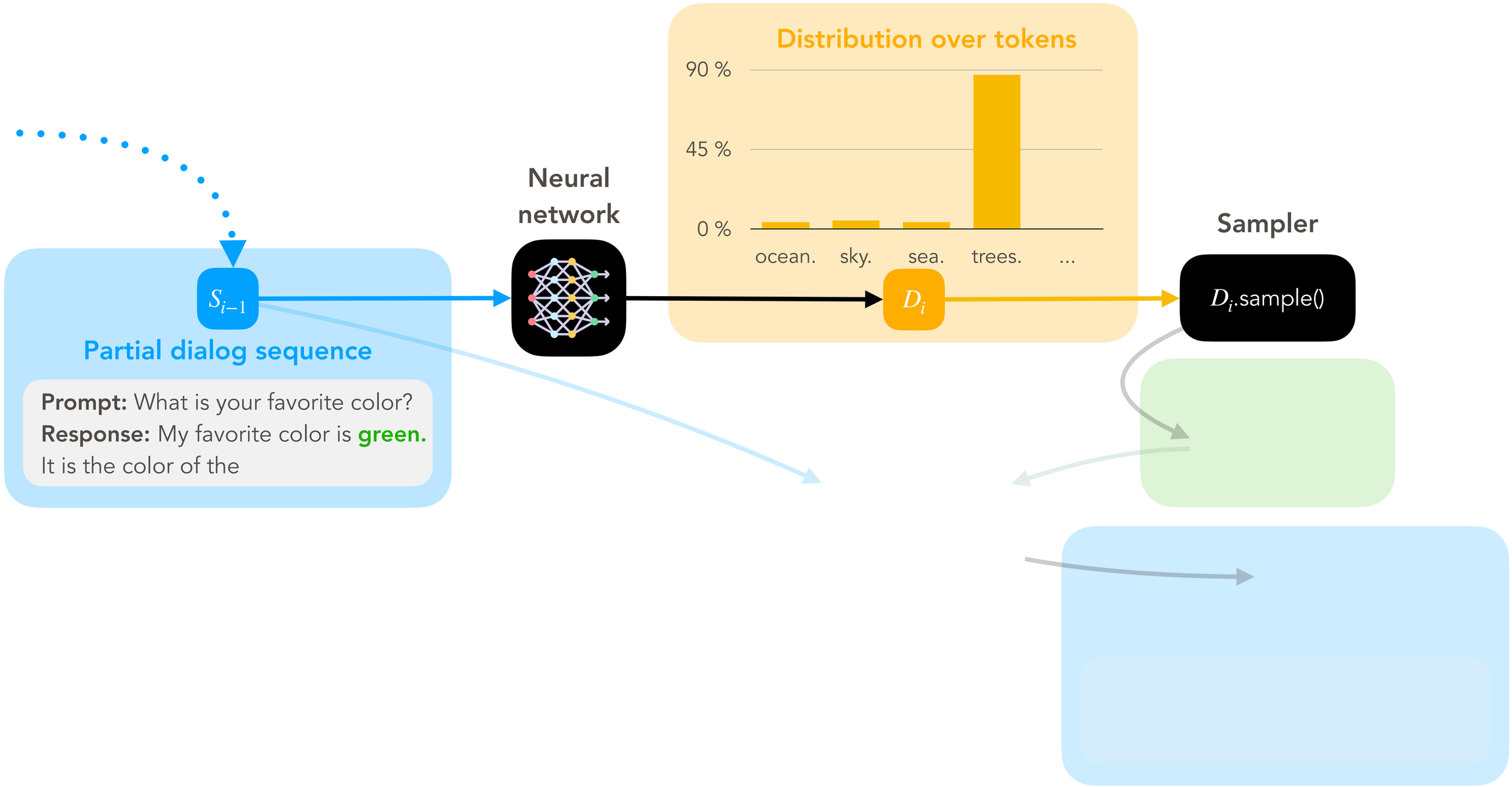
Autoregressive generation through a causal lens



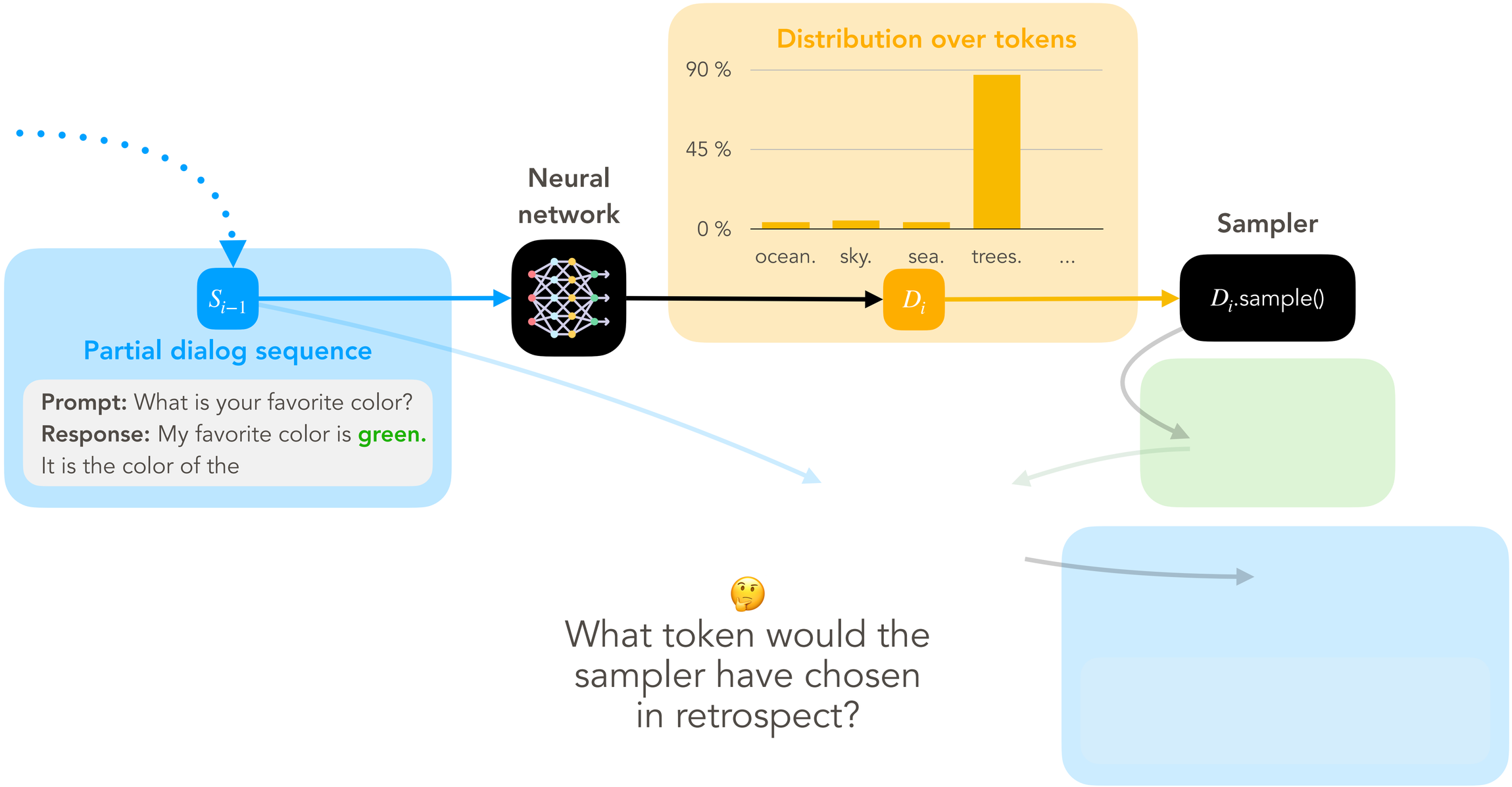
Autoregressive generation through a causal lens



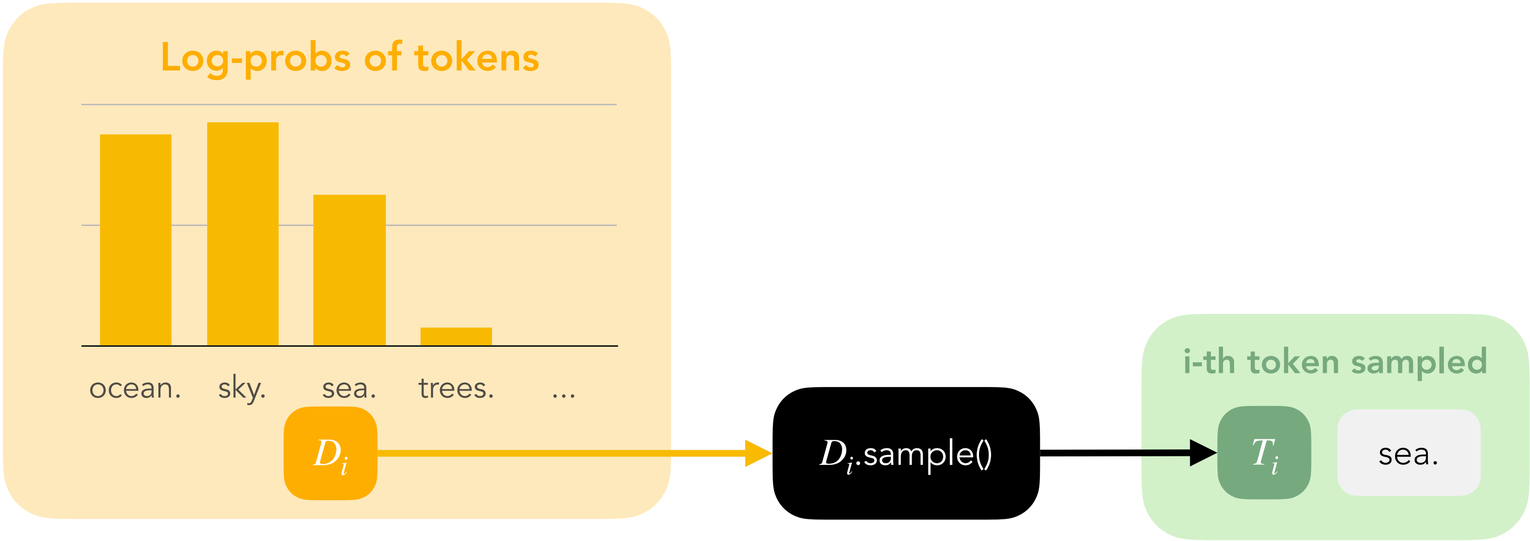
Autoregressive generation through a causal lens



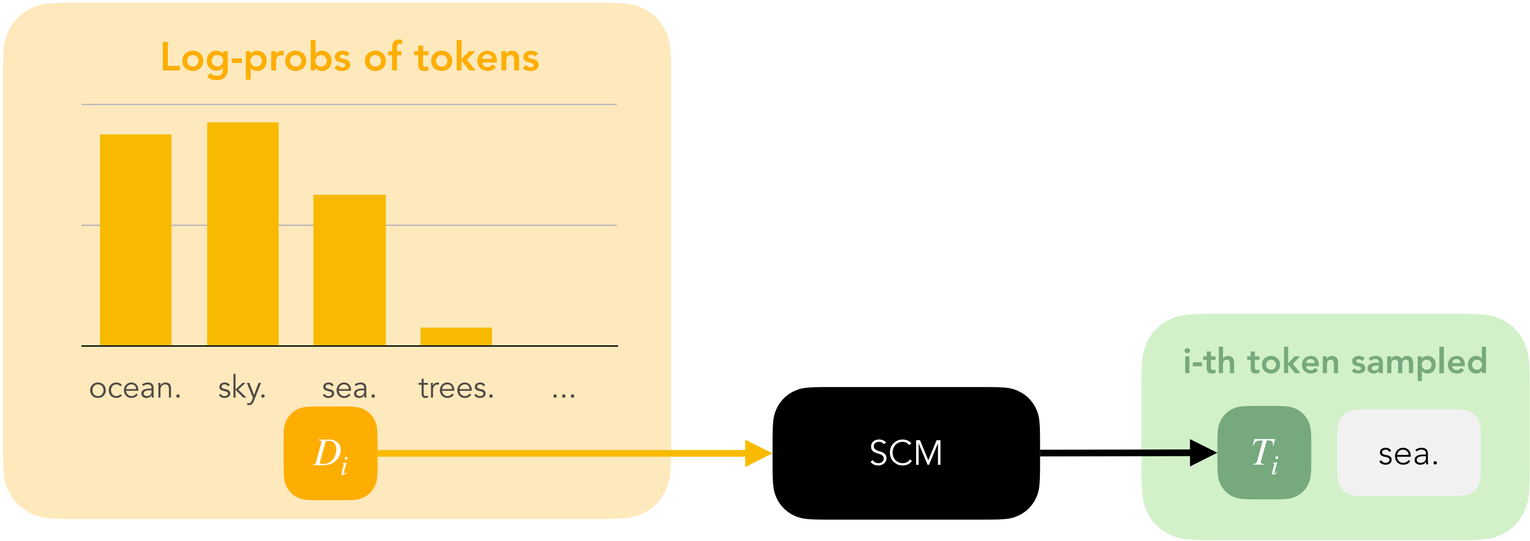
Autoregressive generation through a causal lens



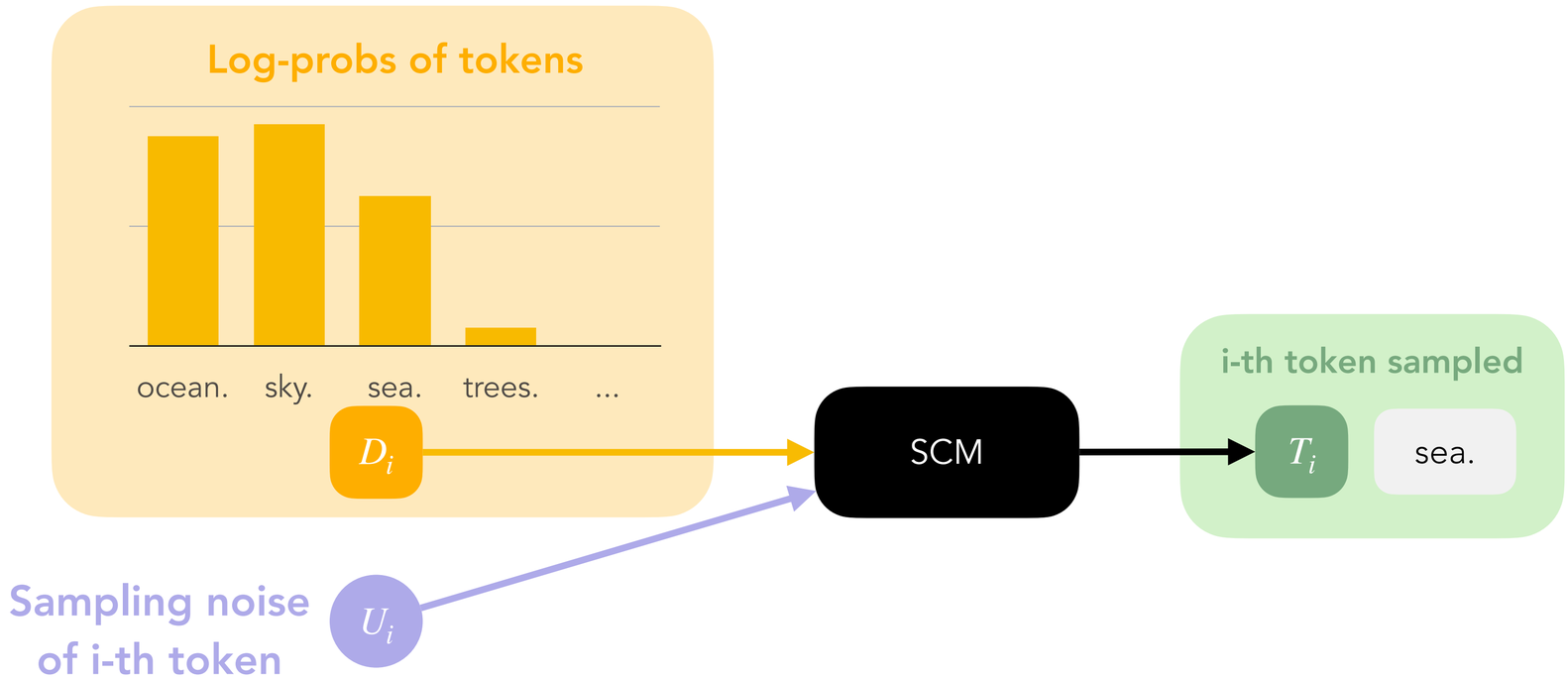
Implementing the sampler as a **Structural Causal Model**



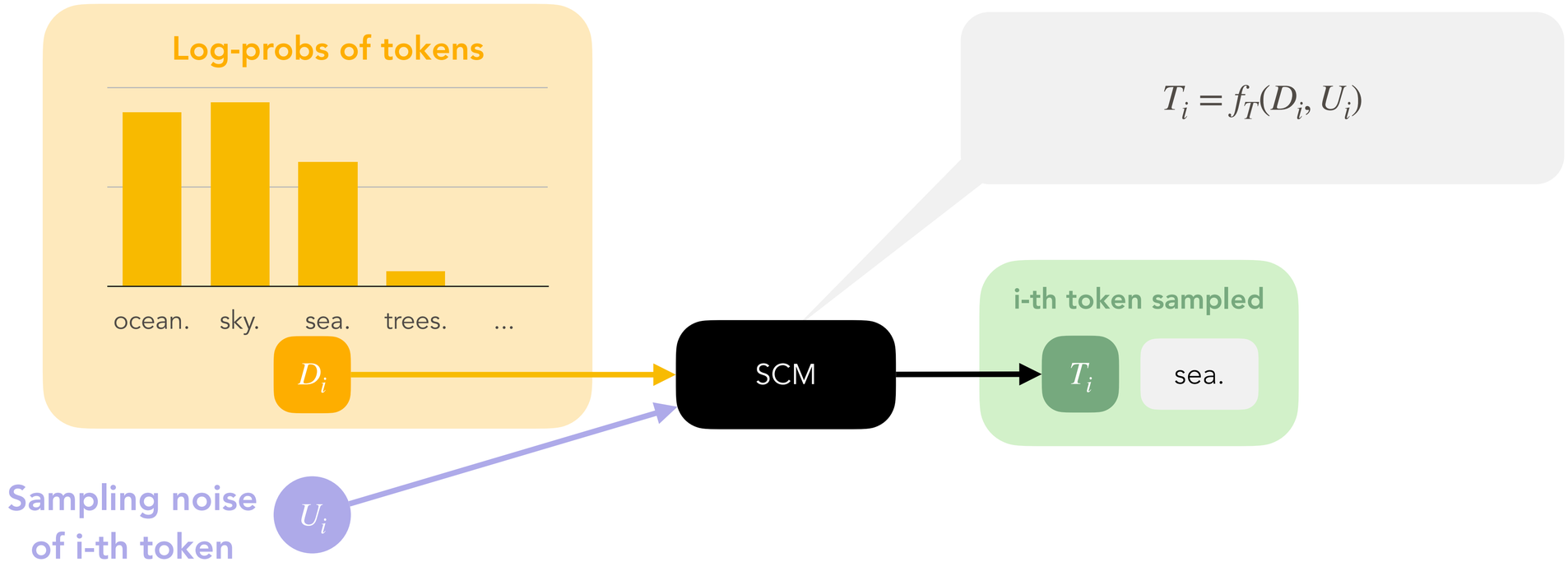
Implementing the sampler as a **Structural Causal Model**



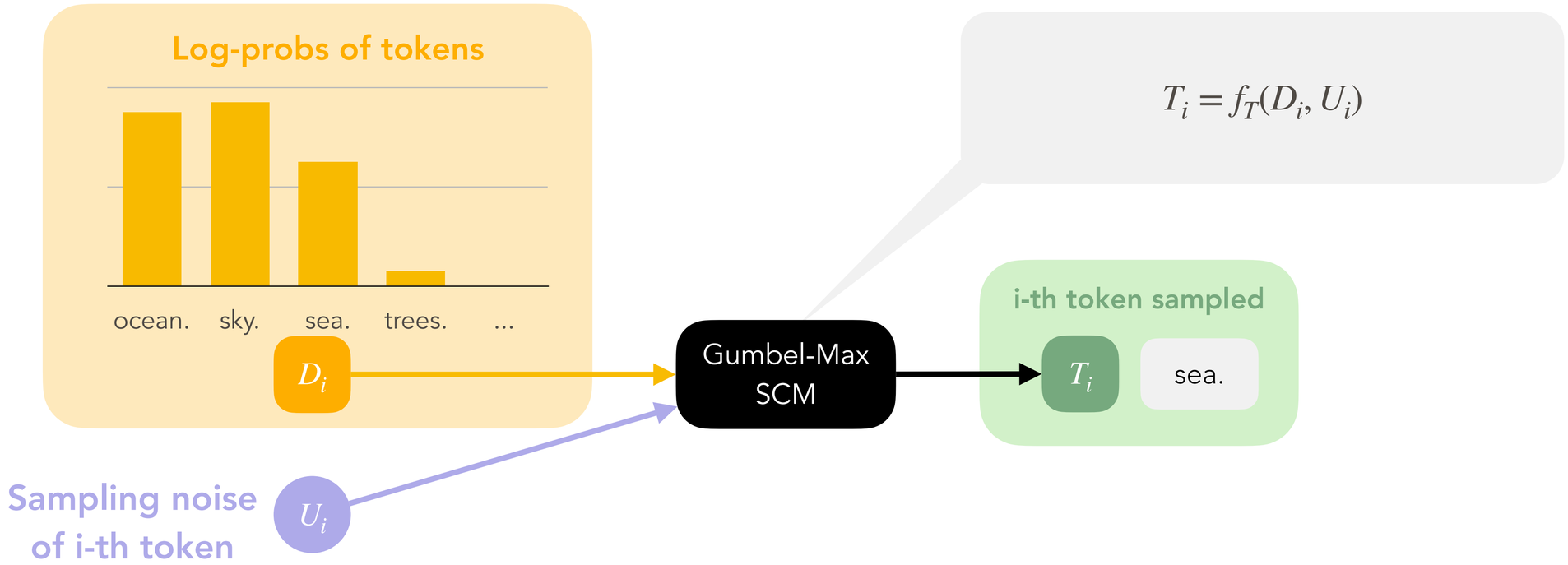
Implementing the sampler as a **Structural Causal Model**



Implementing the sampler as a **Structural Causal Model**

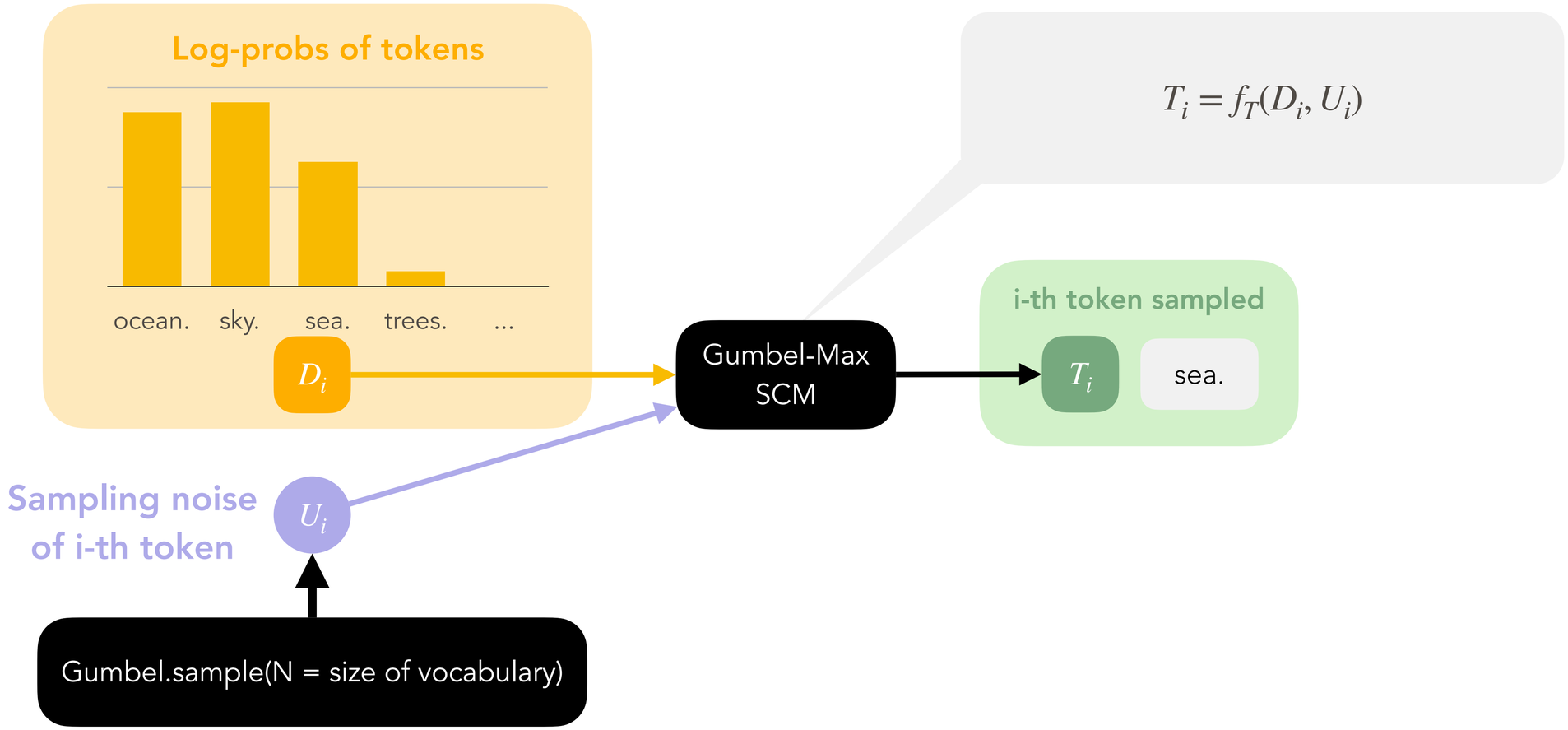


Implementing the sampler as a **Gumbel-max SCM**



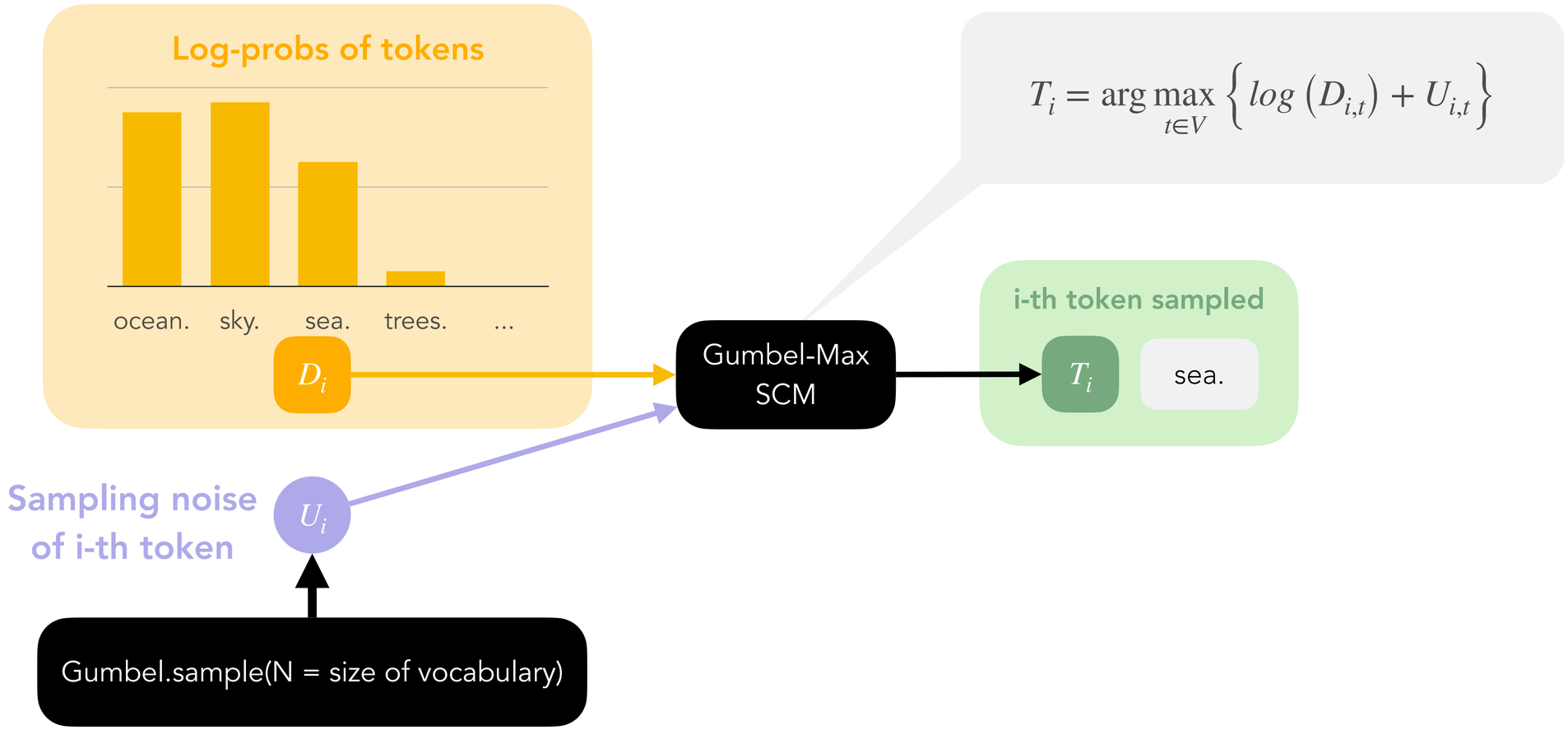
Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Implementing the sampler as a **Gumbel-max SCM**



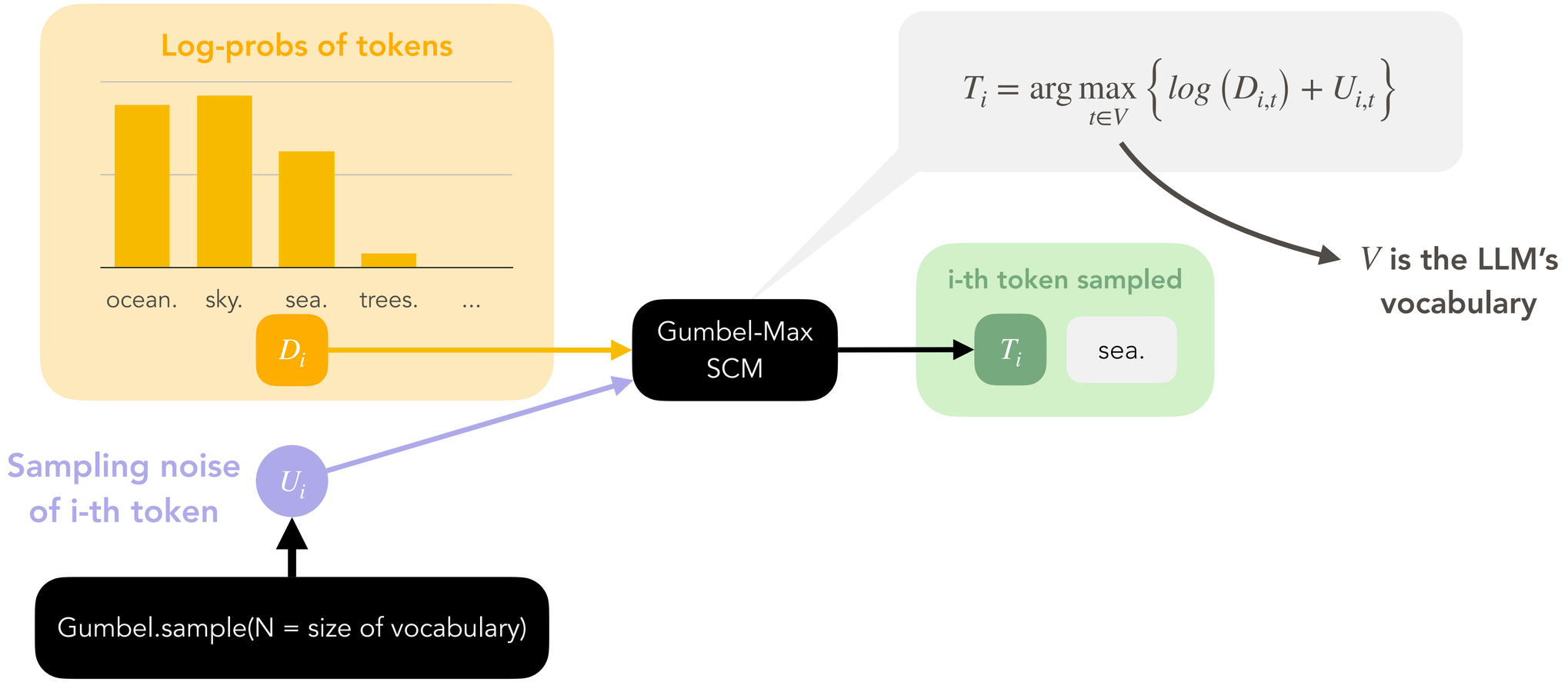
Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Implementing the sampler as a **Gumbel-max SCM**



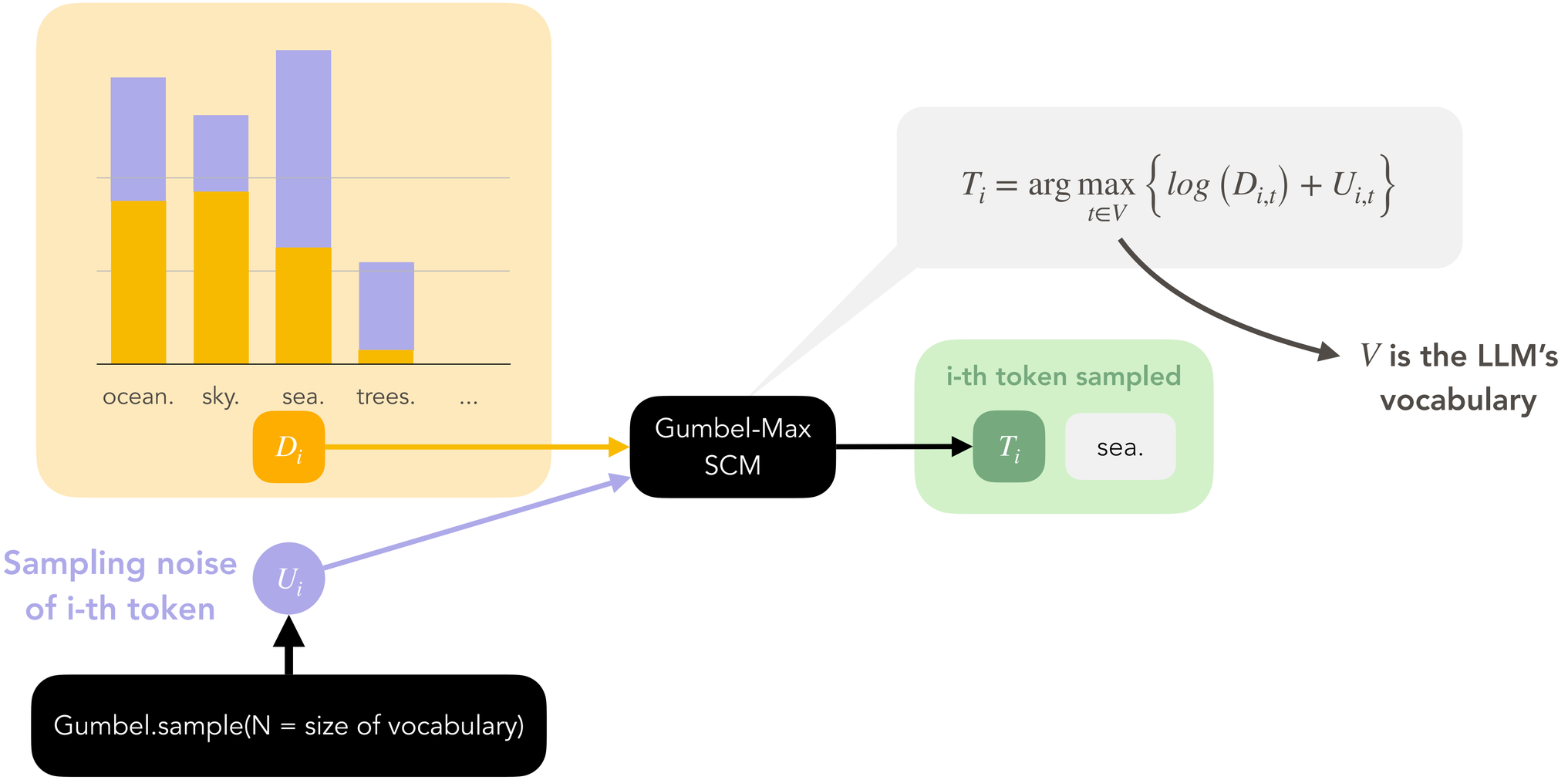
Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Implementing the sampler as a **Gumbel-max SCM**



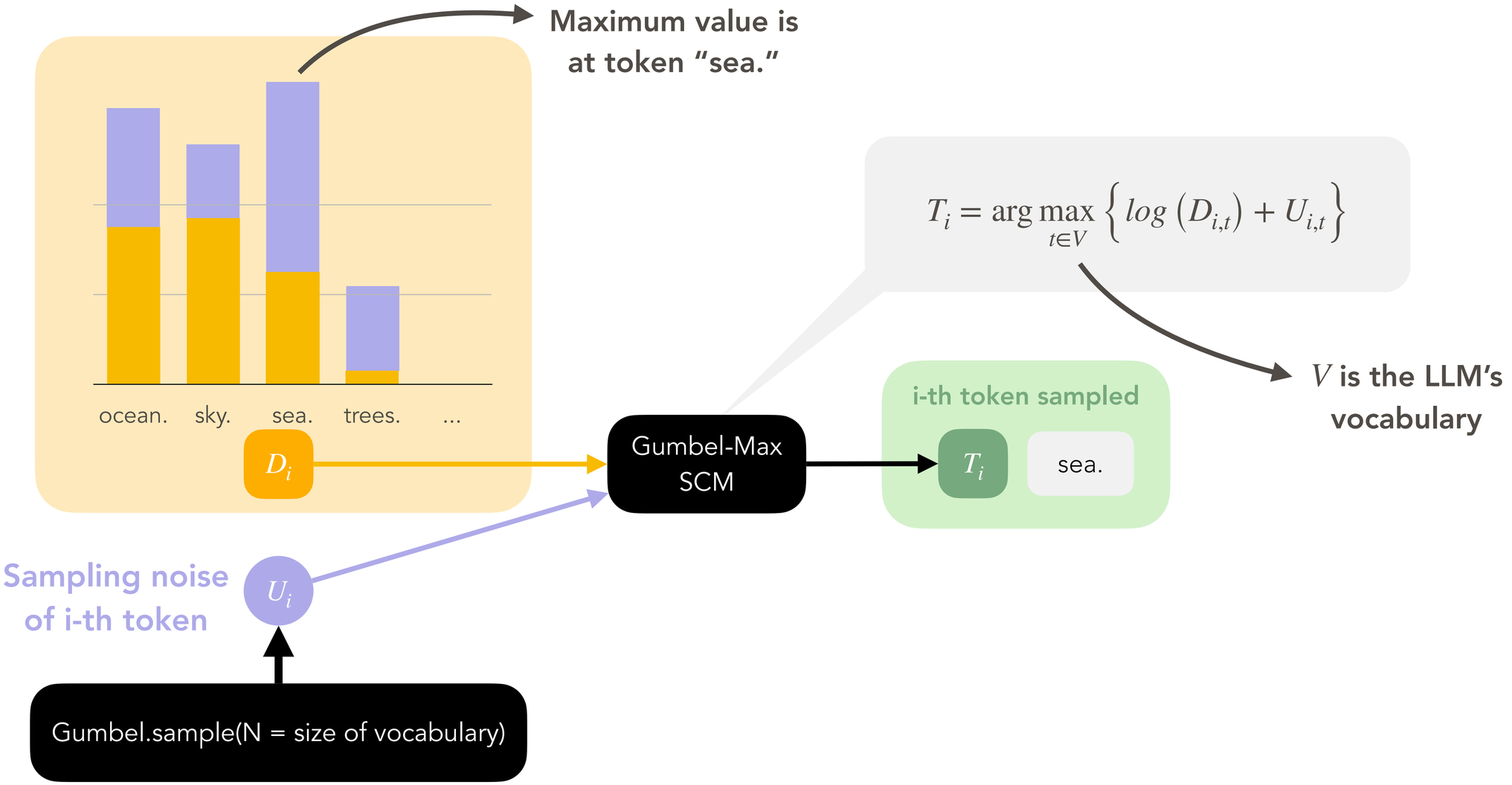
Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Implementing the sampler as a **Gumbel-max SCM**



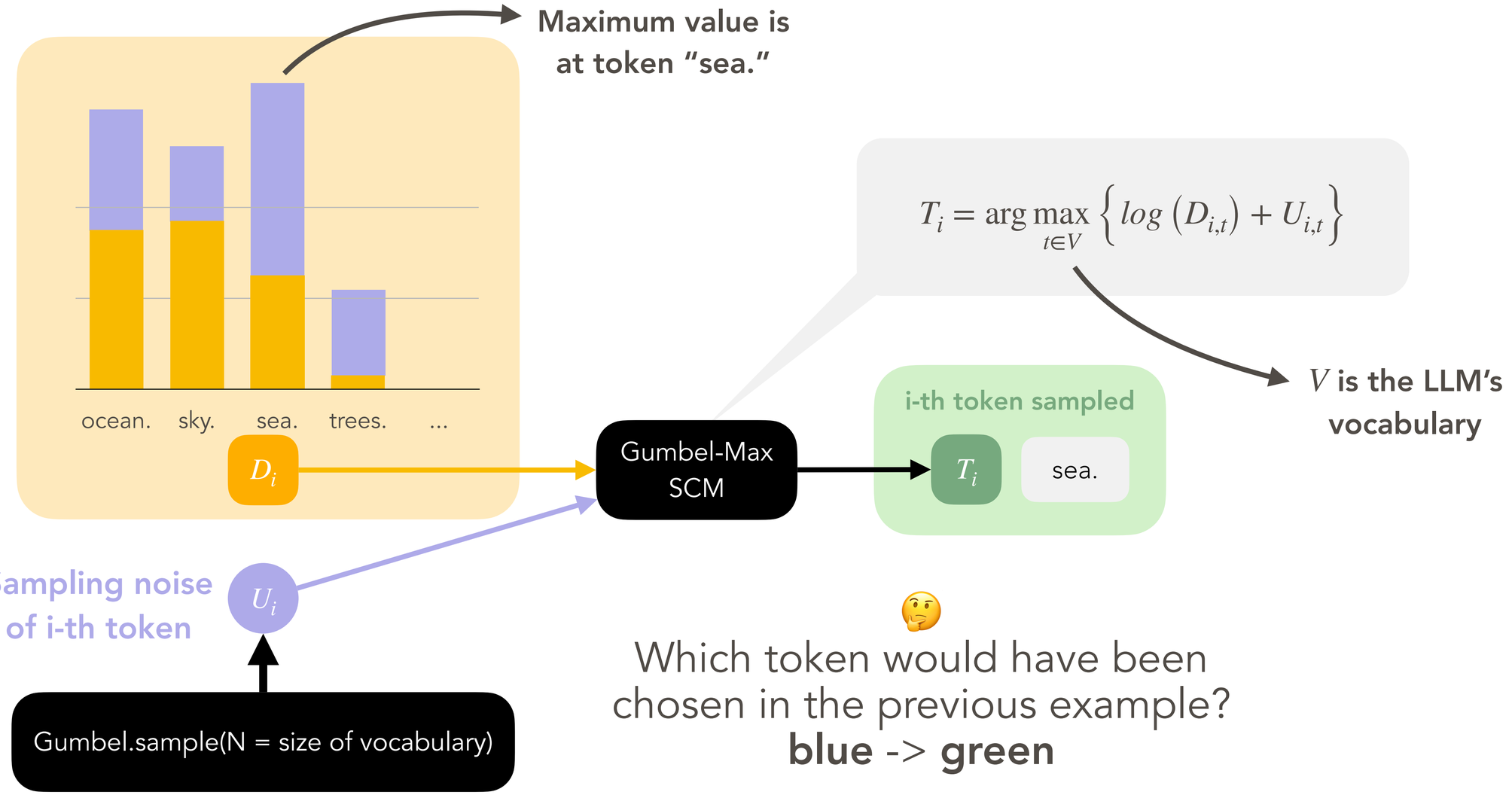
Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Implementing the sampler as a **Gumbel-max SCM**



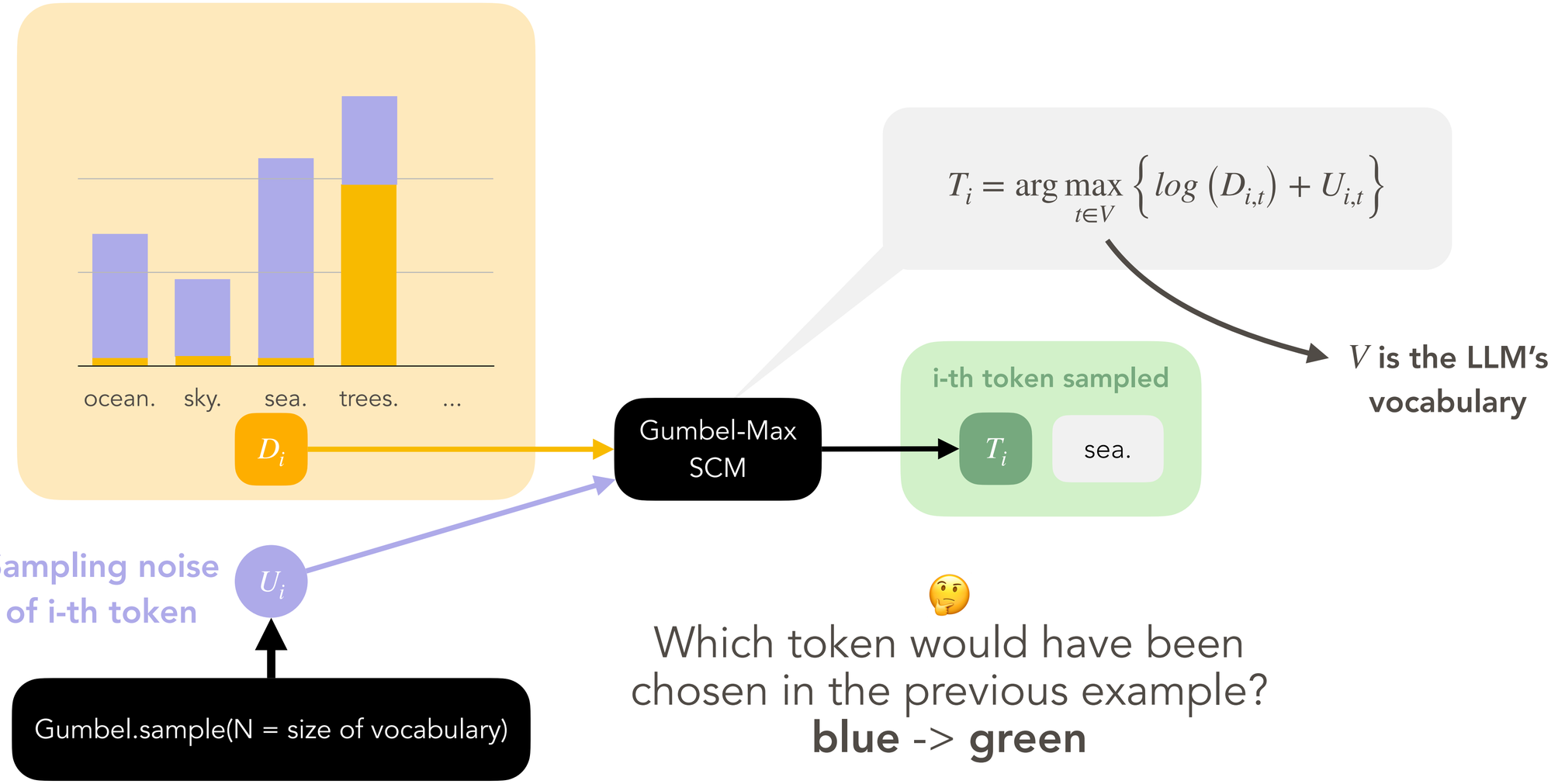
Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Implementing the sampler as a **Gumbel-max SCM**



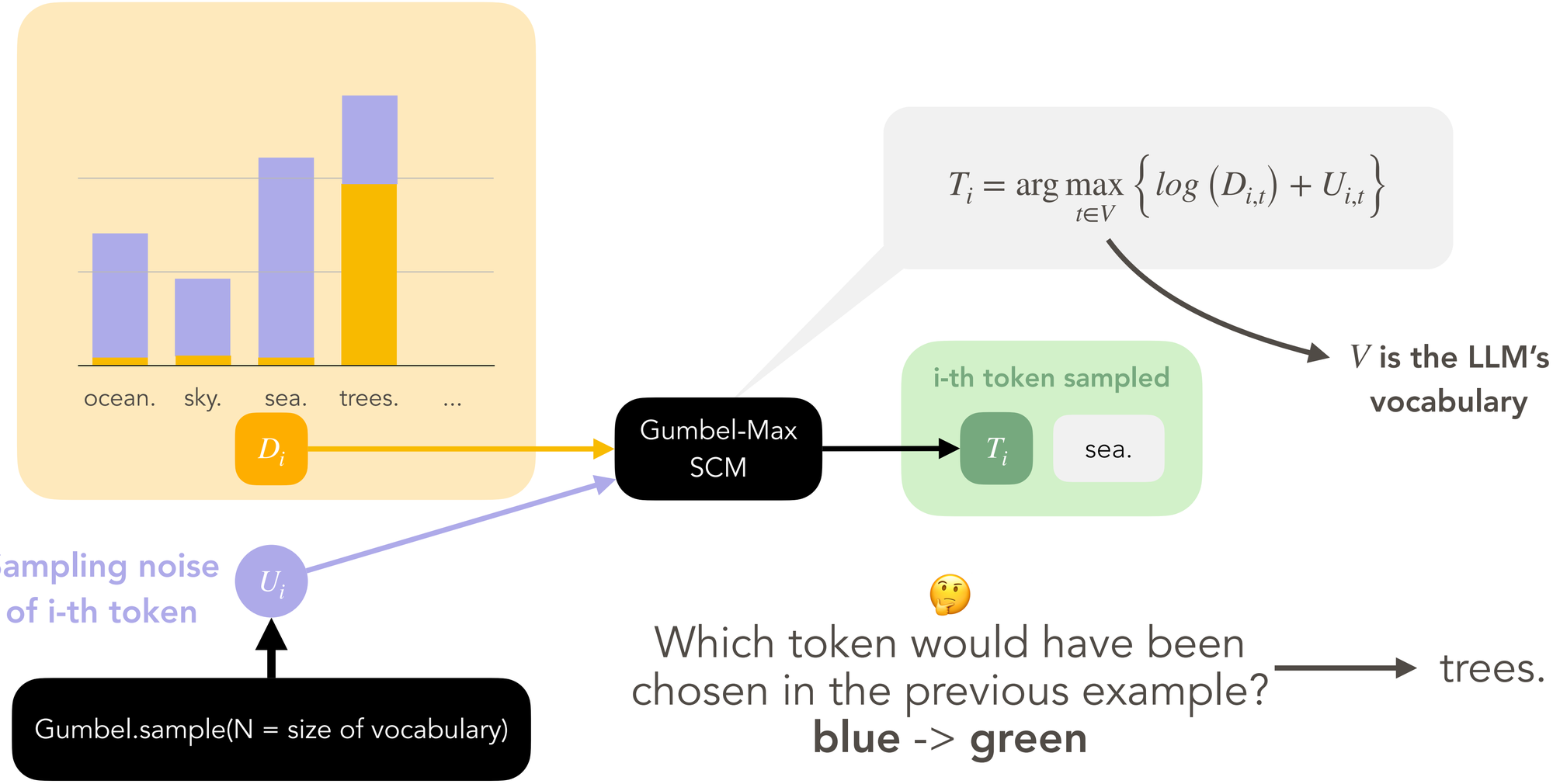
Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Implementing the sampler as a **Gumbel-max SCM**



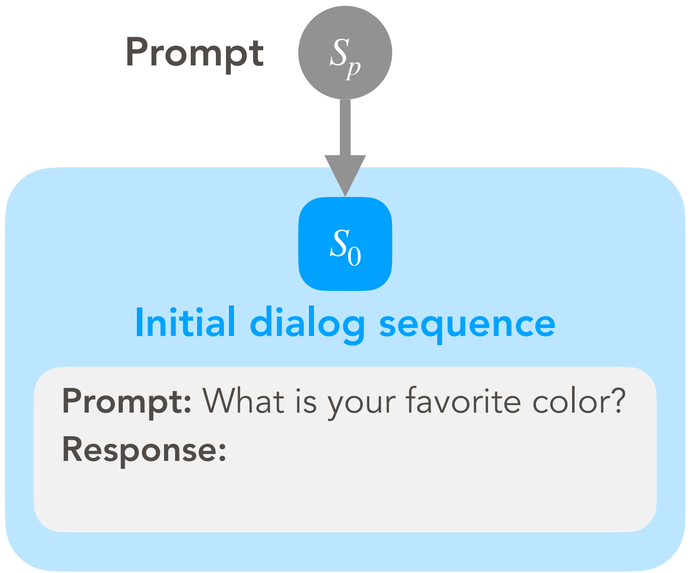
Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Implementing the sampler as a **Gumbel-max SCM**

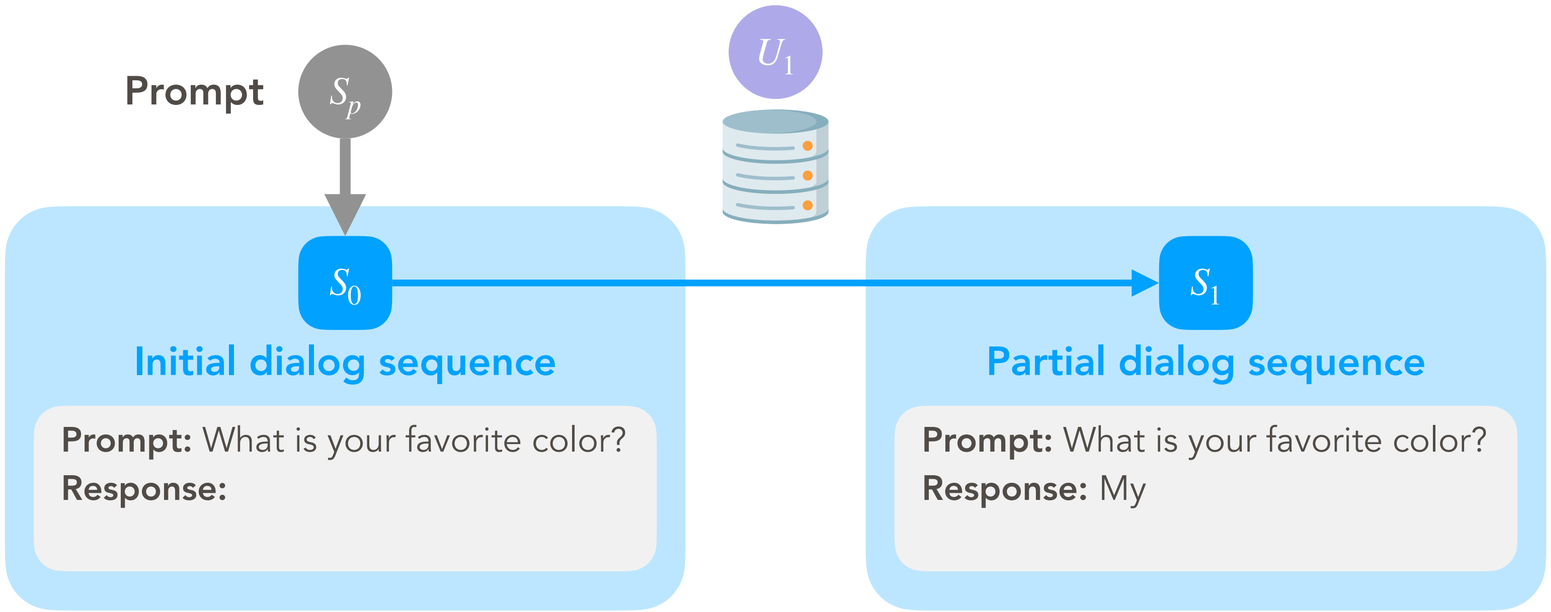


Oberst and Sontag. "*Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*" ICML, 2019.

Finding out what the LLM “would have said”



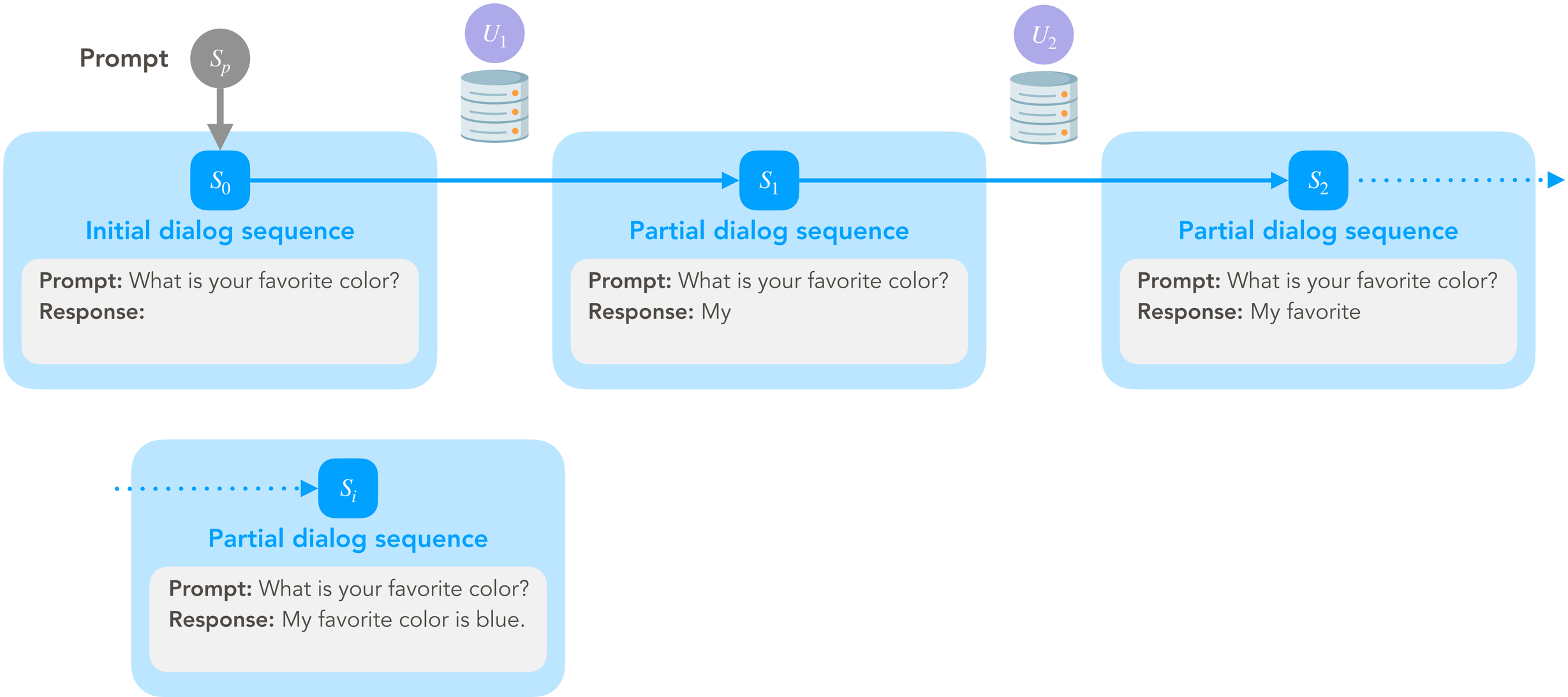
Finding out what the LLM “would have said”



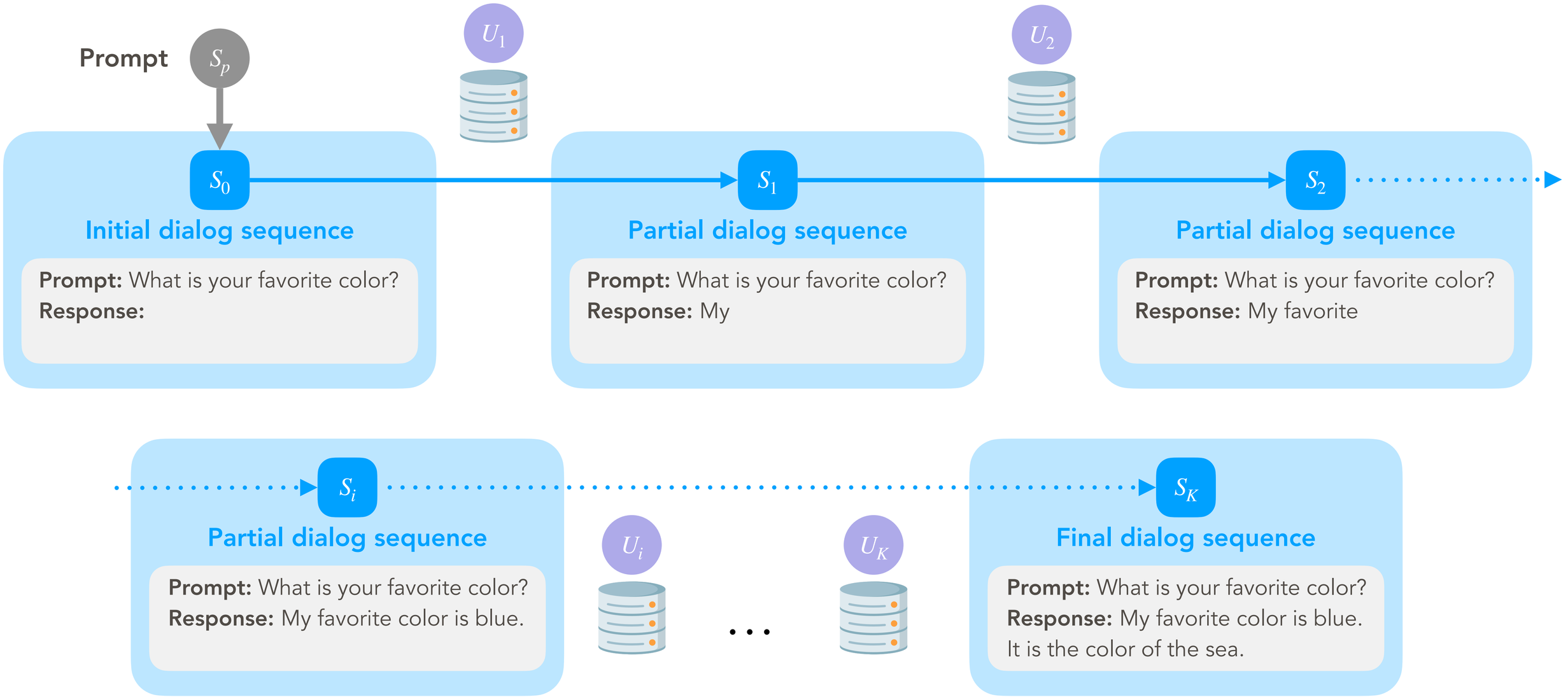
Finding out what the LLM “would have said”



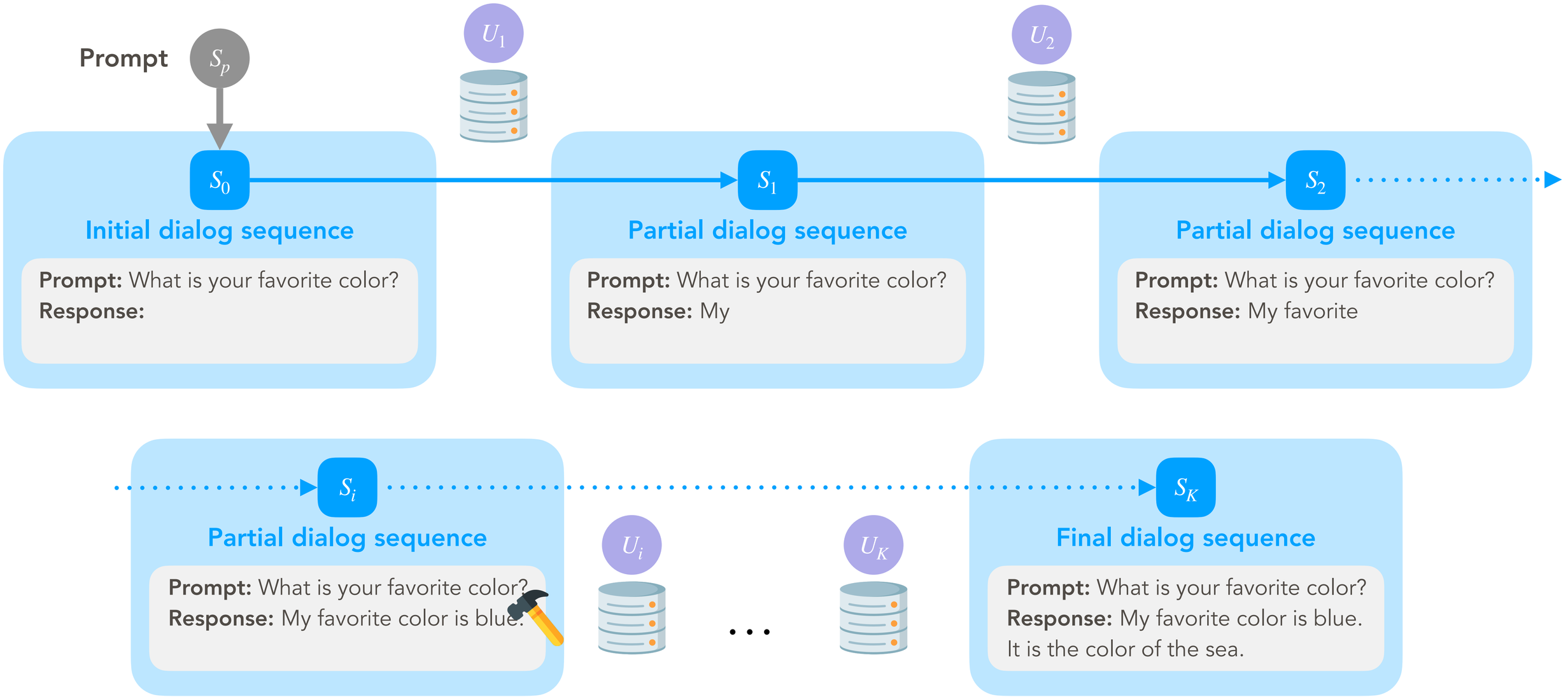
Finding out what the LLM “would have said”



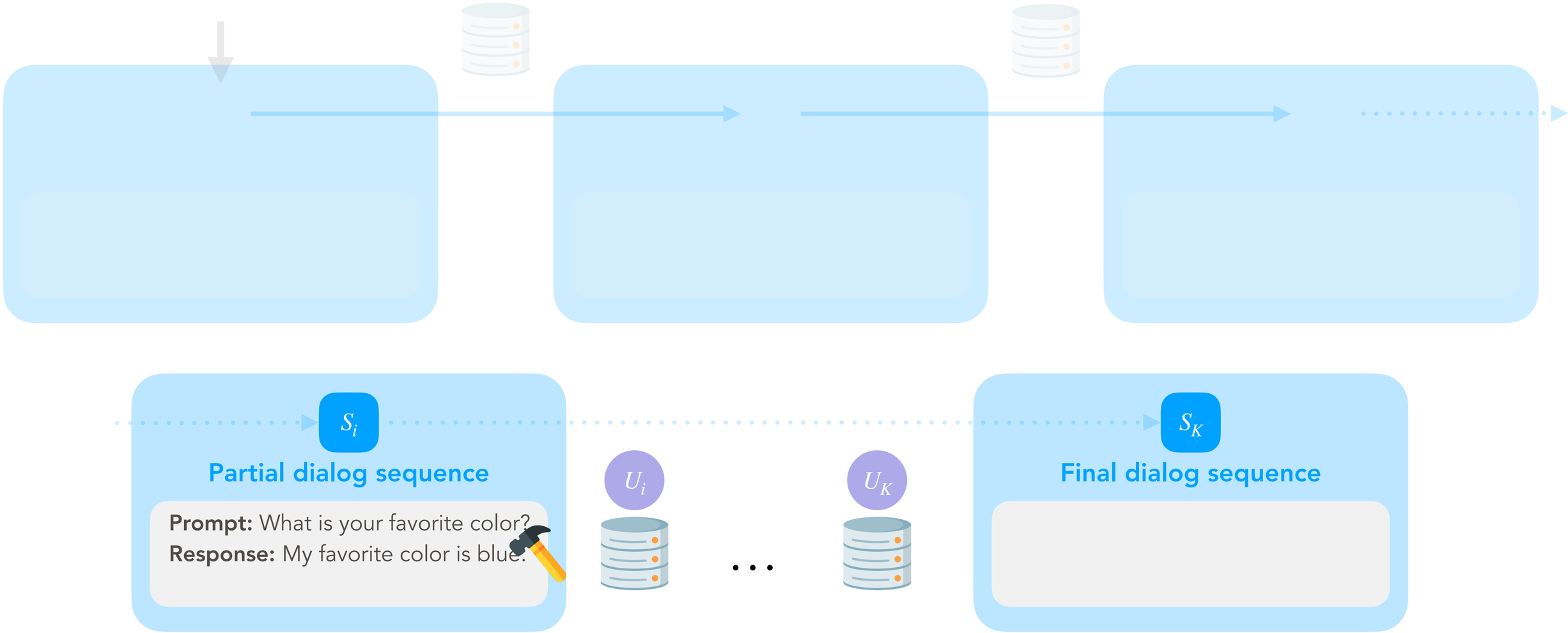
Finding out what the LLM “would have said”



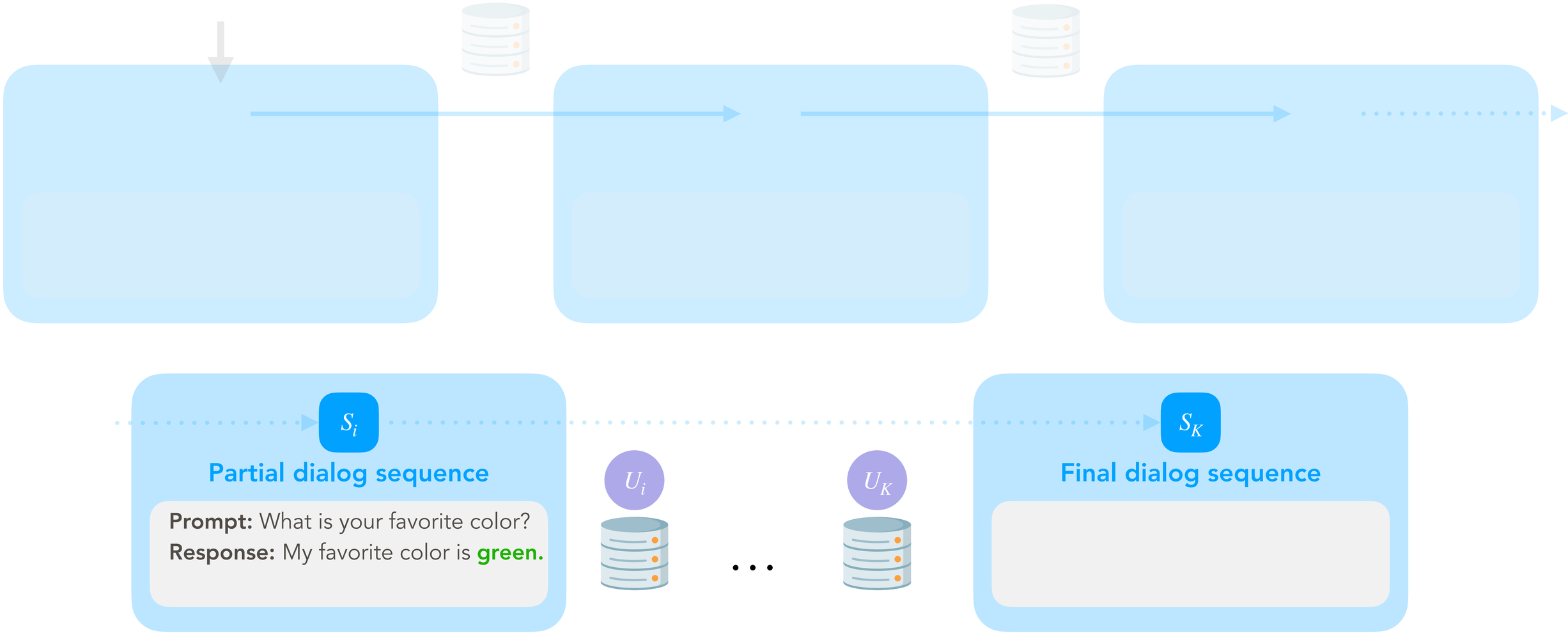
Finding out what the LLM “would have said”



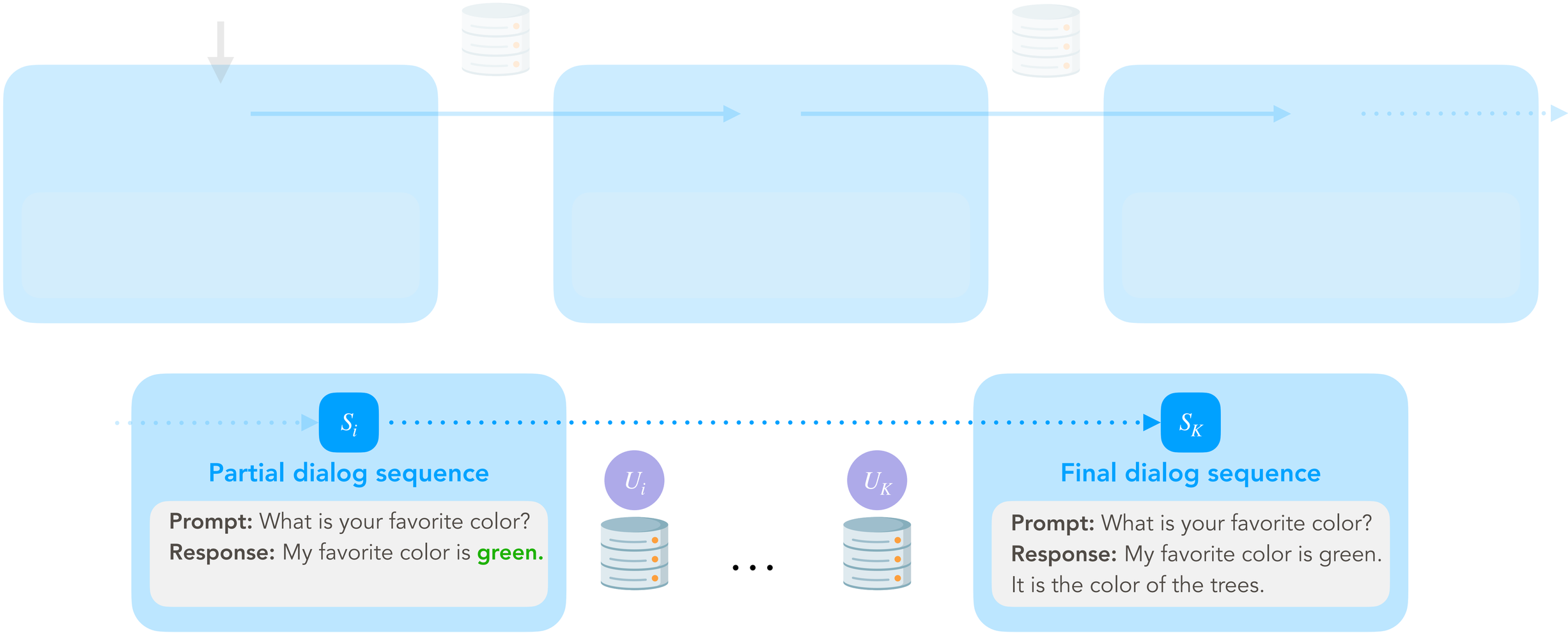
Finding out what the LLM “would have said”



Finding out what the LLM “would have said”



Finding out what the LLM “would have said”



Implementation & experiments



Ministral 8B-Instruct



Llama 3 8B-Instruct

Open-weights models, small, fine-tuned for dialog

- https://docs.mistral.ai/
- https://www.llama.com/

What do counterfactual stories look like?

Prompt: Tell me a fantasy story about a captain. The story should have either a happy or a sad ending.



What do counterfactual stories look like?

Prompt: Tell me a fantasy story about a captain. The story should have either a happy or a sad ending.

Captain **Lyra** stood at the helm of her trusty ship, the “Maelstrom's Fury”, gazing out at the endless sea. For years, she had sailed the seven seas, seeking the legendary Golden Isle, said to grant a single wish to whoever found it. After a lifetime of searching, Lyra finally spotted the island on the horizon. As she stepped foot on its sandy shores, a warm glow enveloped her. Before her, a chest overflowing with glittering gold and jewels lay at her feet. Lyra's heart soared as she reached for the treasure. But, as she touched the gold, the isle began to fade away, leaving her with nothing but a whispered promise: “The true treasure was the journey, not the gold.” Lyra's eyes welled up with tears as she realized the bitter truth — she had sacrificed everything for fleeting riches, and lost the love of her crew, her family, and herself. The “Maelstrom's Fury” sailed away, a lonely shadow on the vast ocean, carrying a captain forever haunted by the what-ifs.

Factual story

What do counterfactual stories look like?

Prompt: Tell me a fantasy story about a captain. The story should have either a happy or a sad ending.

Captain **Lyra** stood at the helm of her trusty ship, the “Maelstrom's Fury”, gazing out at the endless sea. For years, she had sailed the seven seas, seeking the legendary Golden Isle, said to grant a single wish to whoever found it. After a lifetime of searching, Lyra finally spotted the island on the horizon. As she stepped foot on its sandy shores, a warm glow enveloped her. Before her, a chest overflowing with glittering gold and jewels lay at her feet. Lyra's heart soared as she reached for the treasure. But, as she touched the gold, the isle began to fade away, leaving her with nothing but a whispered promise: “The true treasure was the journey, not the gold.” Lyra's eyes welled up with tears as she realized the bitter truth — she had sacrificed everything for fleeting riches, and lost the love of her crew, her family, and herself. The “Maelstrom's Fury” sailed away, a lonely shadow on the vast ocean, carrying a captain forever haunted by the what-ifs.

Factual story

Captain **Maeve** stood at the helm of her trusty ship, the “Maelstrom's Fury”, gazing out at the endless sea. For years, she had sailed the seven seas, seeking the legendary Golden Isle, said to grant a single wish to whoever found it. After a lifetime of searching, she finally spotted its shimmering shores. As she stepped onto the island, a radiant light enveloped her, and she wished for her lost love, the brave sailor, Eamon, to return to her. Suddenly, Eamon emerged from the mist, frail but alive. Overjoyed, Maeve scooped him into her arms, and they shared a tender kiss as the sun dipped into the horizon, their love reunited at last.

Counterfactual story



Do the LLMs show signs of bias?

- We ask LLMs to generate fictional census data
- Each individual has a list of features:
  - Age
  - Sex
  - Citizenship
  - Race
  - …
  - Income
  - Education

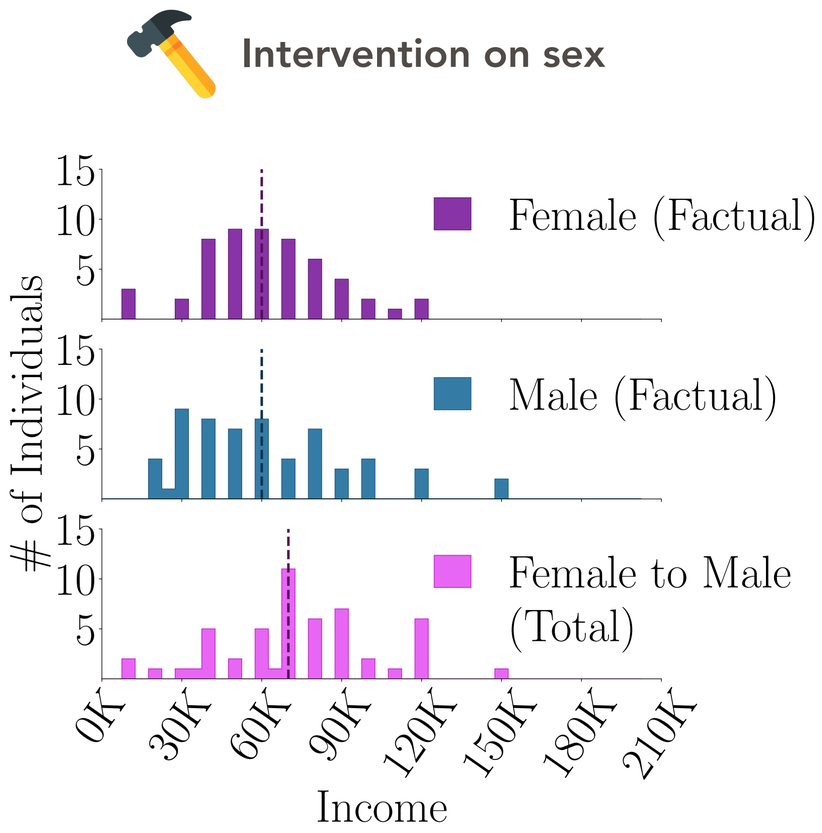


Intervention on sex



Do the LLMs show signs of bias?

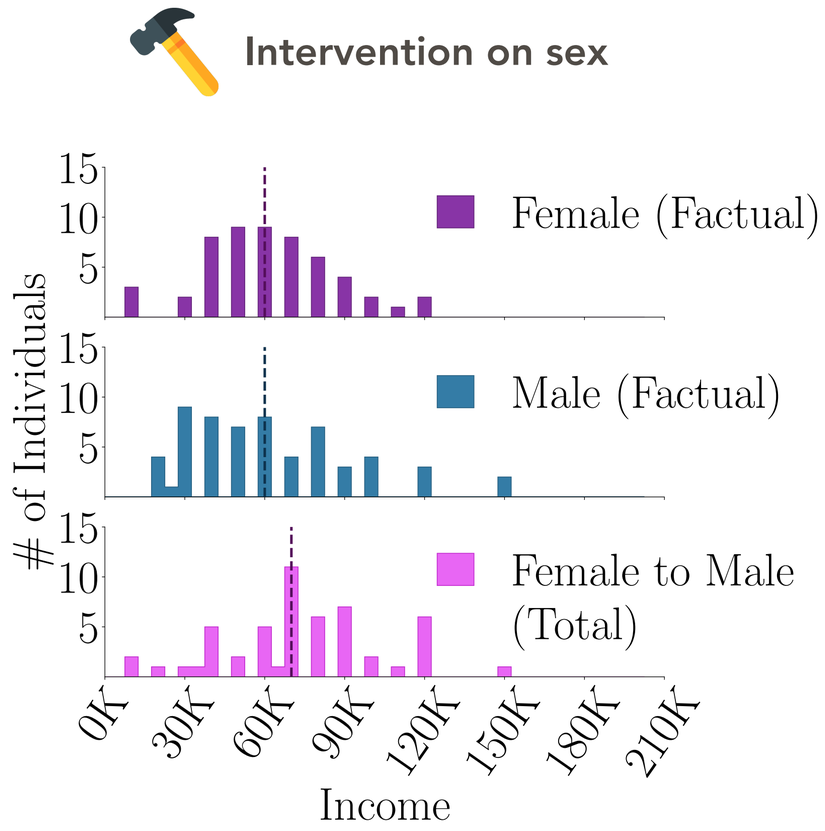
- We ask LLMs to generate fictional census data
- Each individual has a list of features:
  - Age
  - Sex
  - Citizenship
  - Race
  - …
  - Income
  - Education



Do the LLMs show signs of bias?

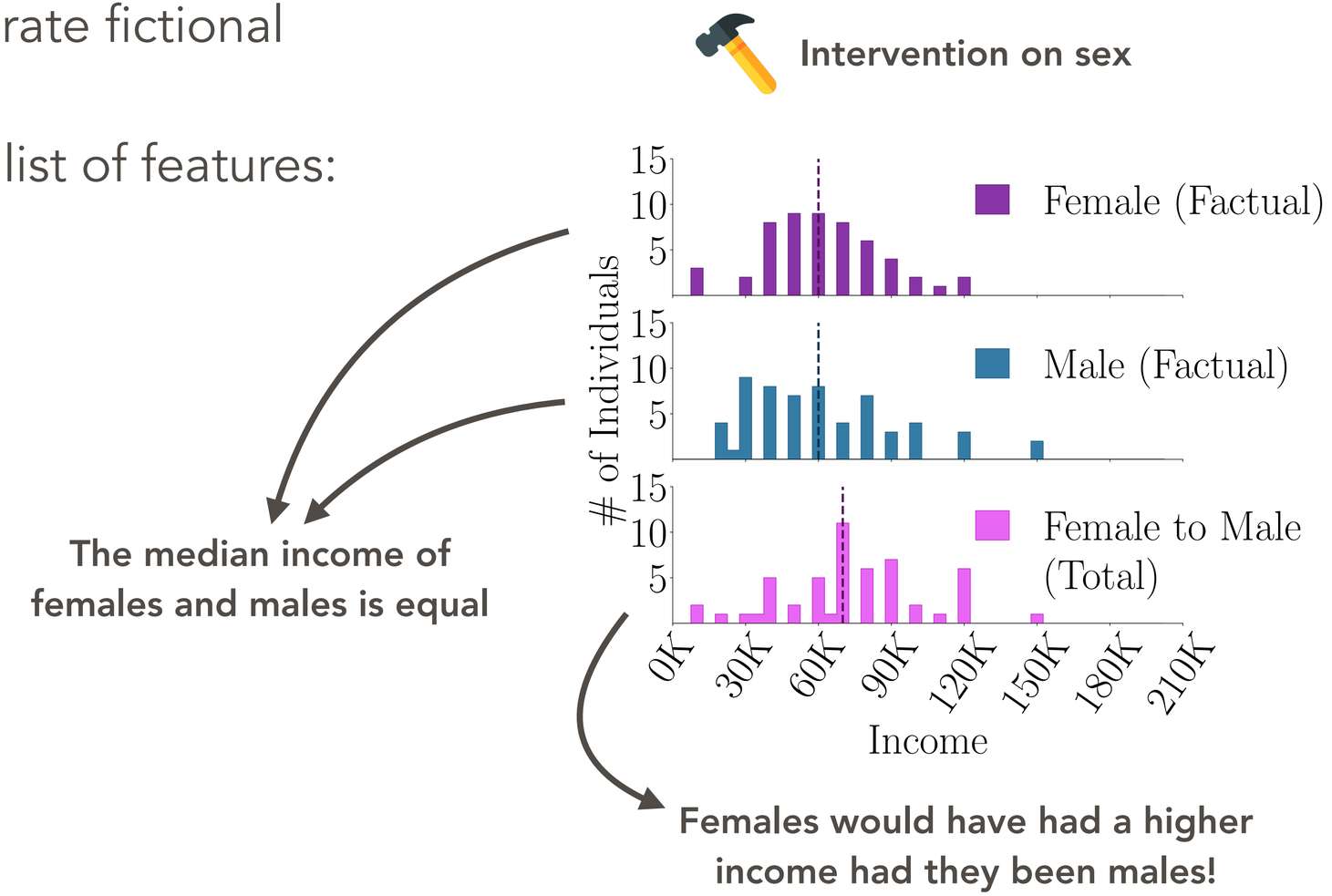
- We ask LLMs to generate fictional census data
- Each individual has a list of features:
  - Age
  - Sex
  - Citizenship
  - Race
  - …
  - Income
  - Education

The median income of females and males is equal



Do the LLMs show signs of bias?

- We ask LLMs to generate fictional census data
- Each individual has a list of features:
  - Age
  - Sex
  - Citizenship
  - Race
  - …
  - Income
  - Education





Code



Demo

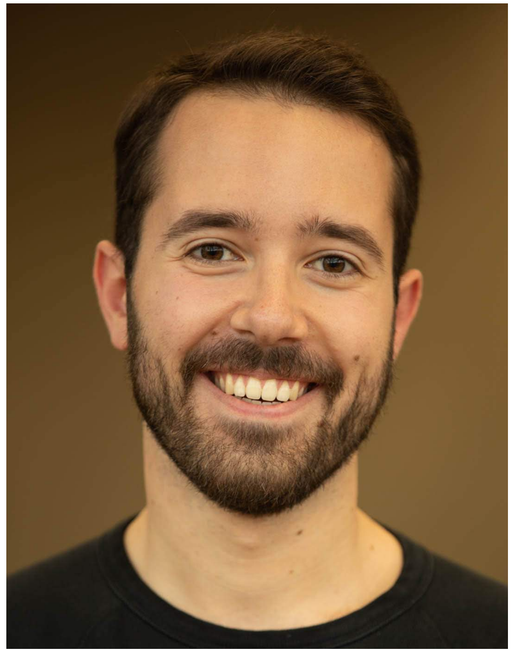
Thanks for your attention!



Tobias Gerstenberg
gerstenberg@stanford.edu



Manuel Gomez-Rodriguez
manuel@mpi-sws.org



Stratis Tsirtsis
stsirtsis@mpi-sws.org