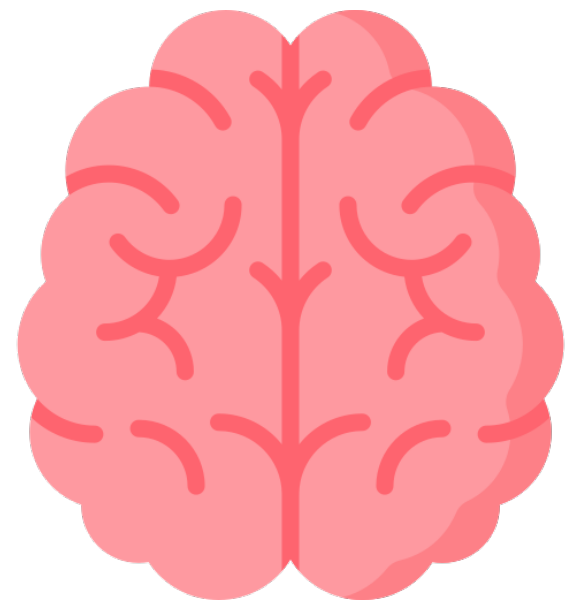
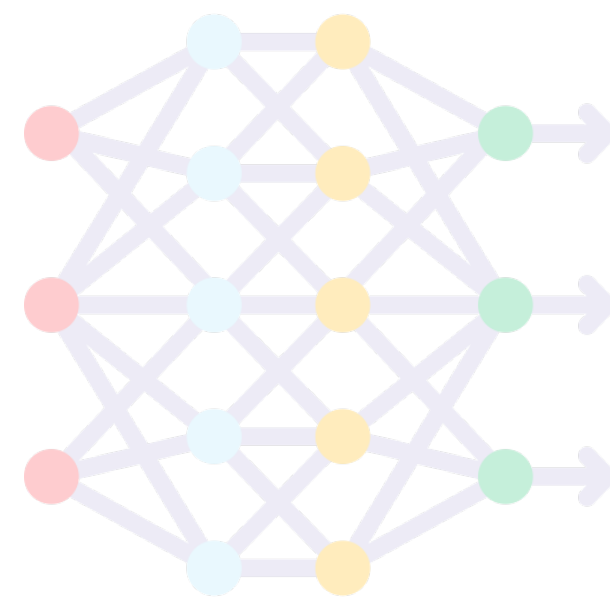


Outline



Cognitive
science



Machine
learning

AaI

Large language
models

Could I have done anything better?



Could I have done anything better?



Analysis

1

2

3

4

5

6

7

8

mdclermont (1226)

4... Nc6	591	40%	39%	21%
4... Be7	29	24%	38%	38%
4... Be6	10	50%	40%	10%
4... g6	4	100%		
4... h6	2	100%		
4... a6	1	100%		
4... b6	1	100%		
4... Nf6	1	100%		

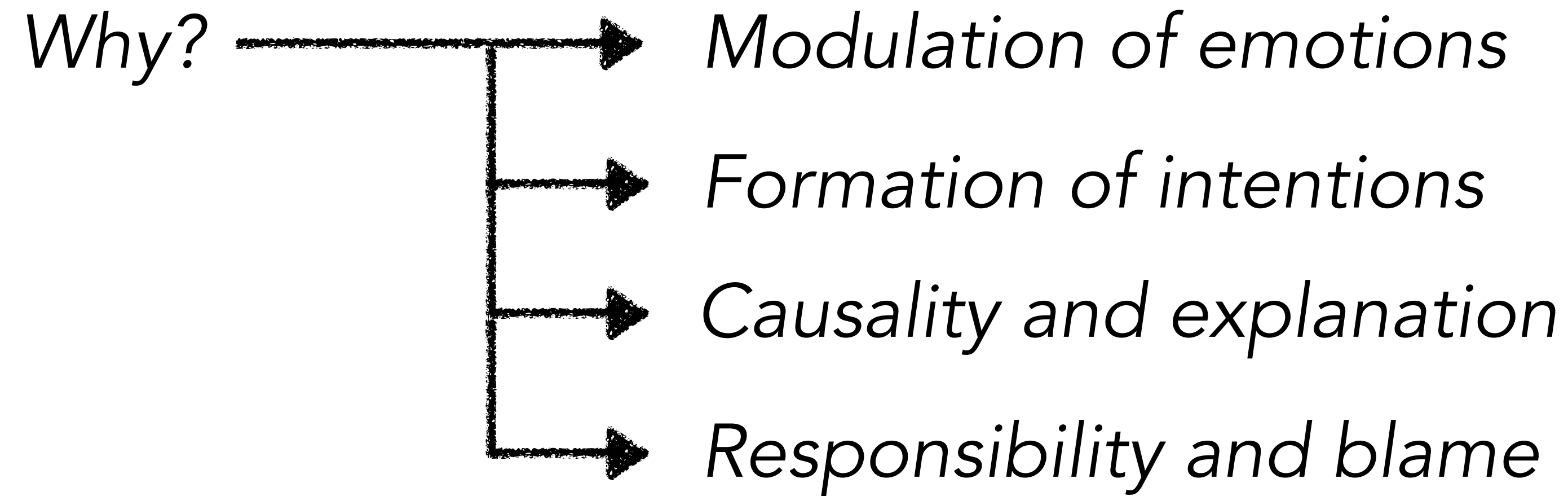
McCurds (931)

We think of counterfactuals all the time

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

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

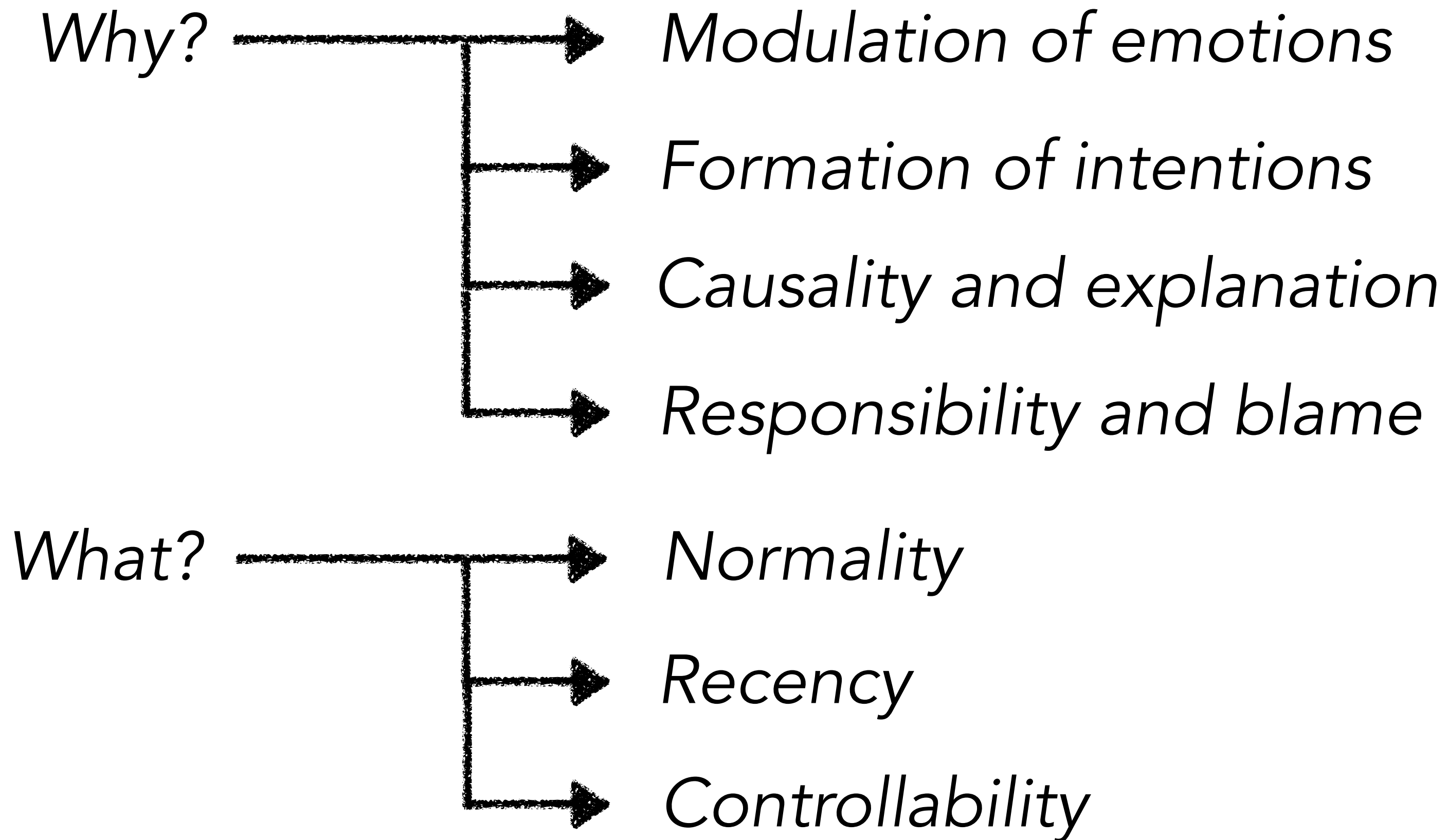
We think of counterfactuals all the time



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

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

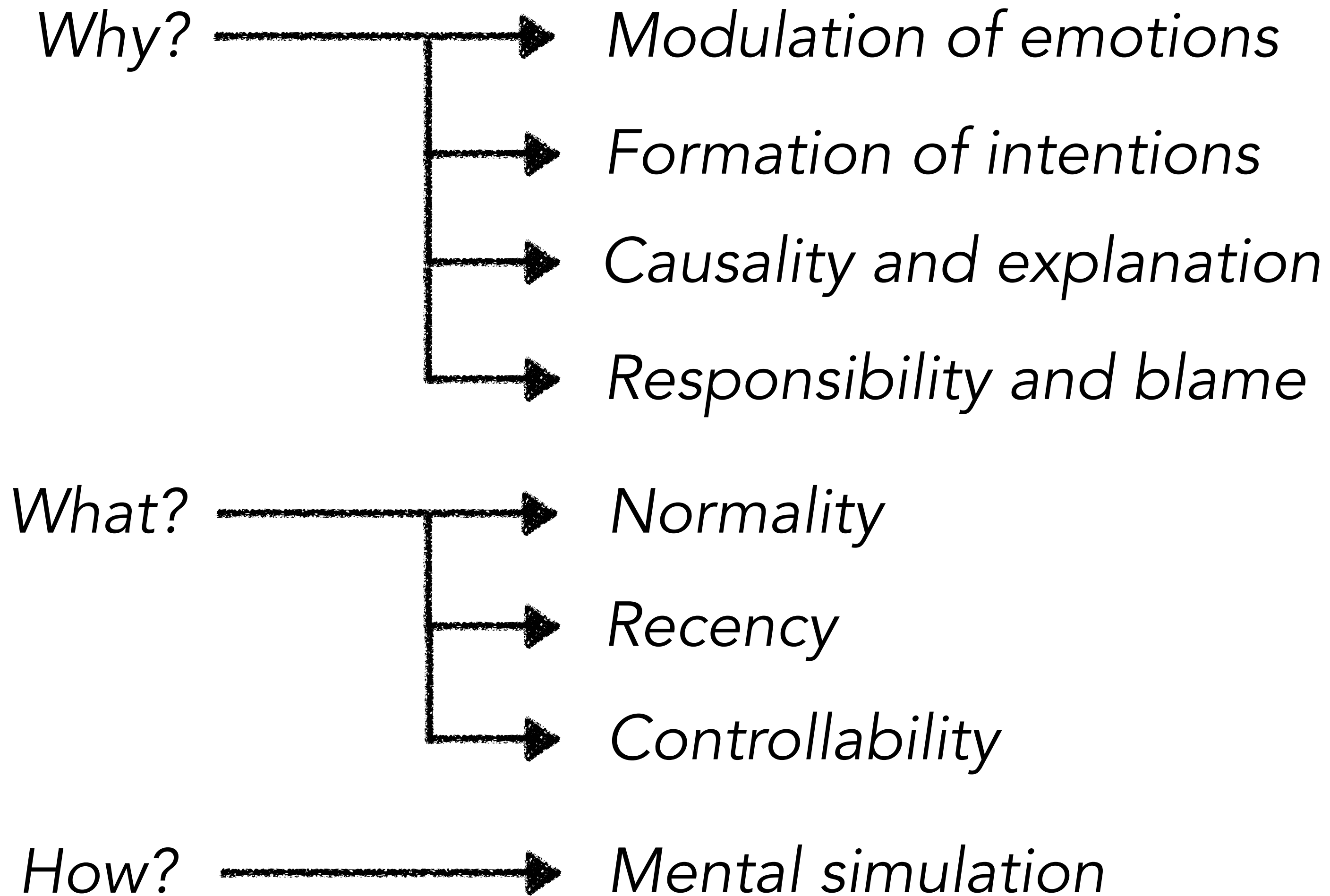
We think of counterfactuals all the time



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

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

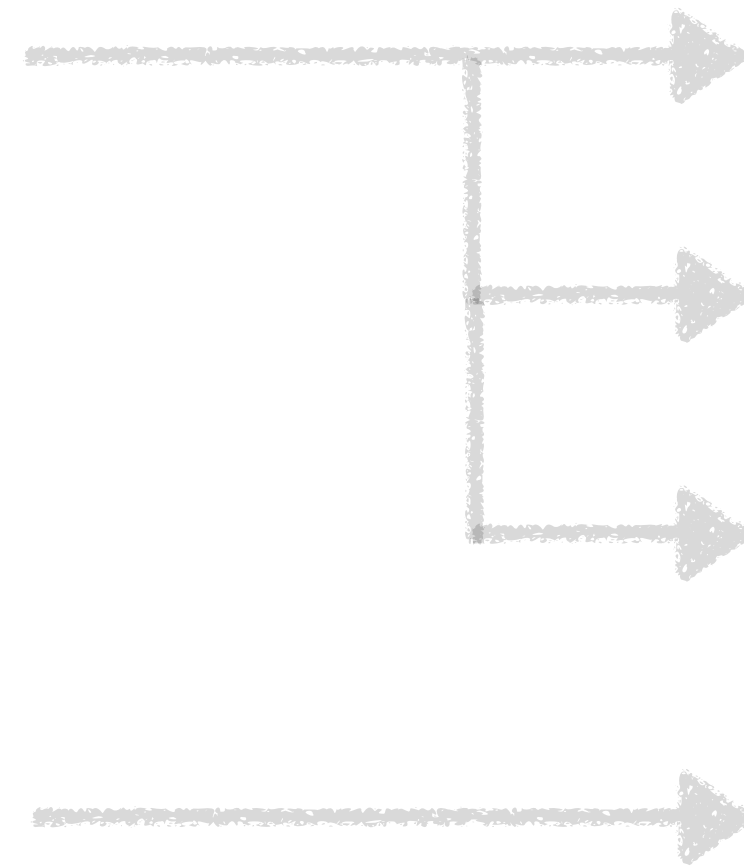
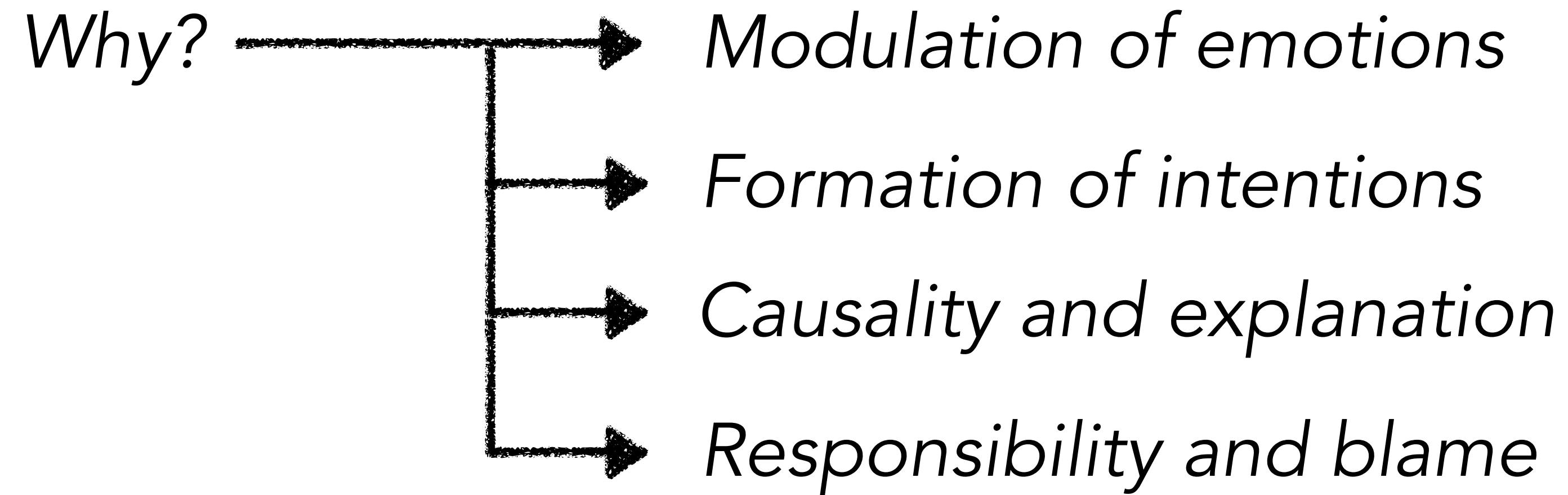
We think of counterfactuals all the time



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

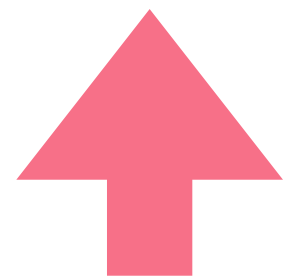
Following a (typically) bad event, we tend to think in terms of counterfactuals that could have led to a **better** or **worse** outcome.



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.



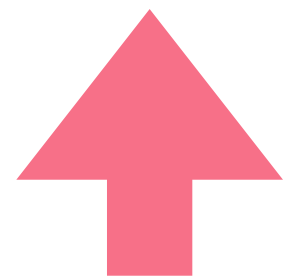
I could have won



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

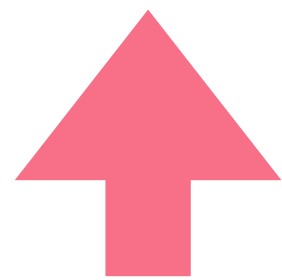
Downward counterfactuals lead to positive emotions



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

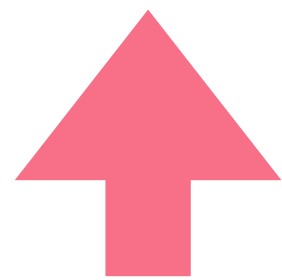
Upward counterfactuals lead to negative emotions



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

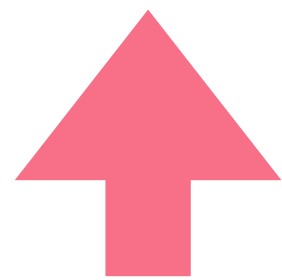


I could have lost the medal

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

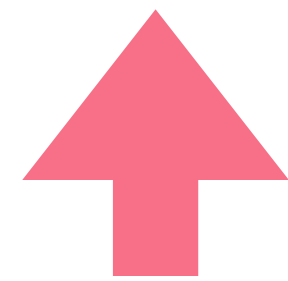
Formation of intentions

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.

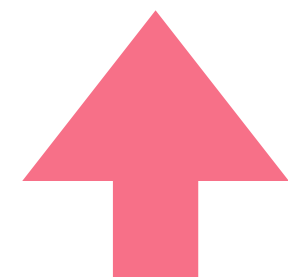


If I had understood the controller's words, I
wouldn't have initiated the landing attempt

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

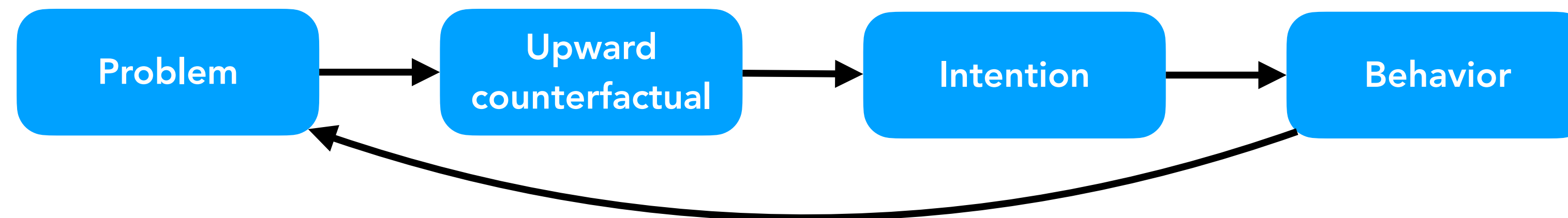
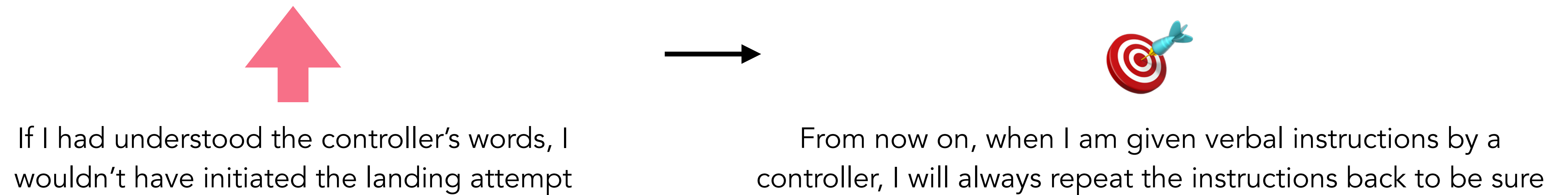


From now on, when I am given verbal instructions by a controller, I will always repeat the instructions back to be sure

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.

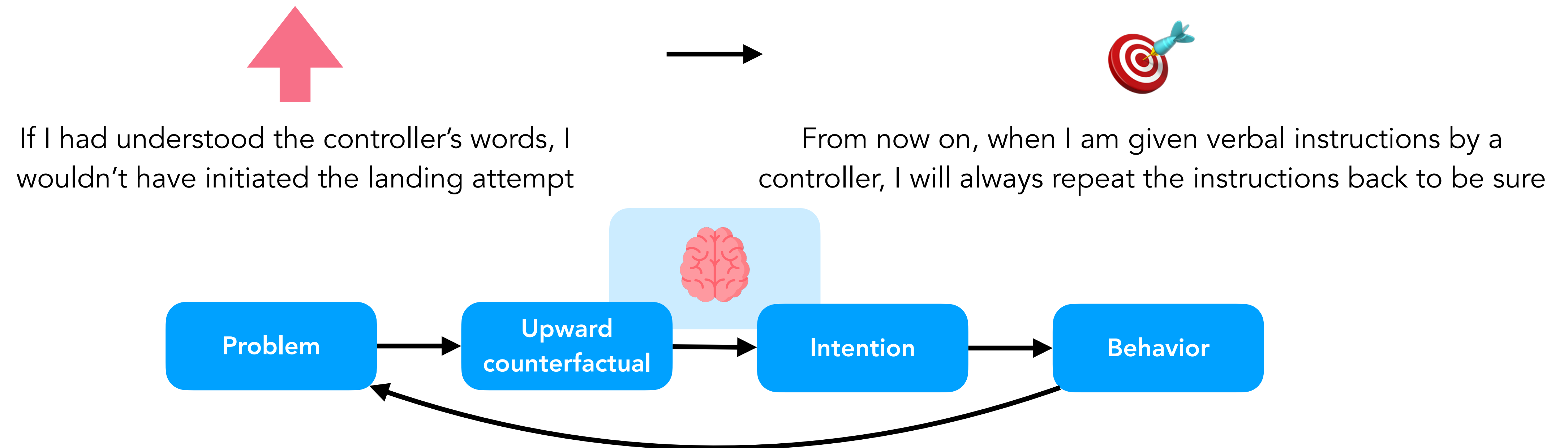


Morris & Moore. *"The lessons we (don't) learn: counterfactual thinking and organizational accountability after a close call."* Administrative Science Quarterly, 2000.

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.



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.

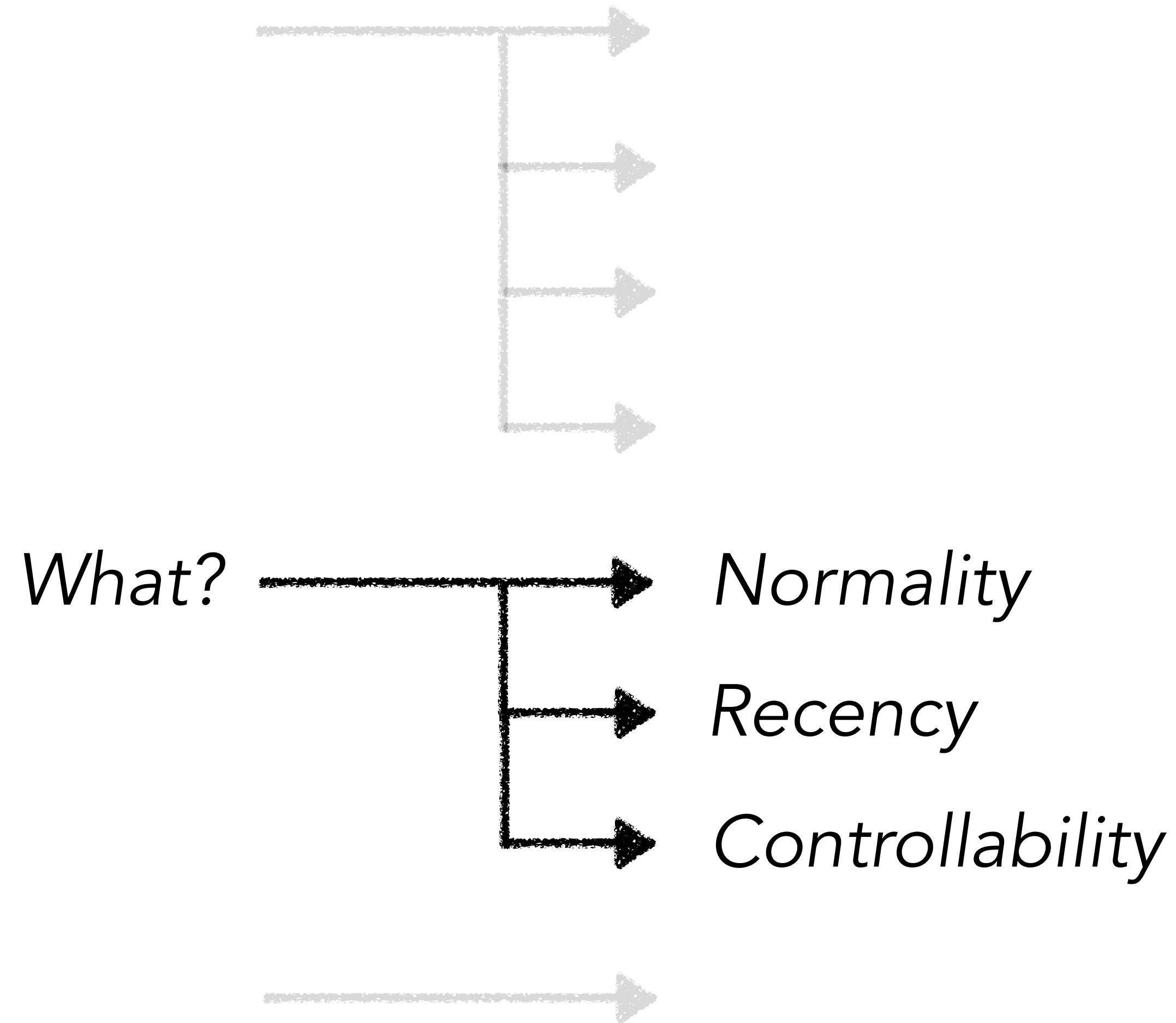
Hart and Honoré. "Causation in the Law". Oxford University Press, 1985.

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

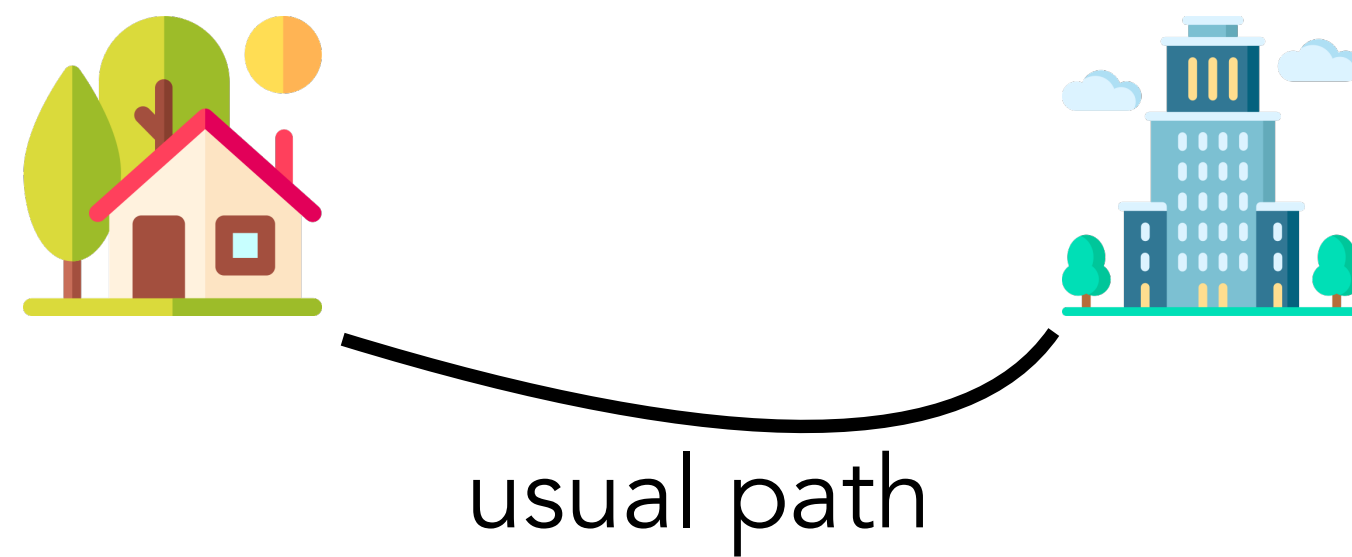
Normality



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

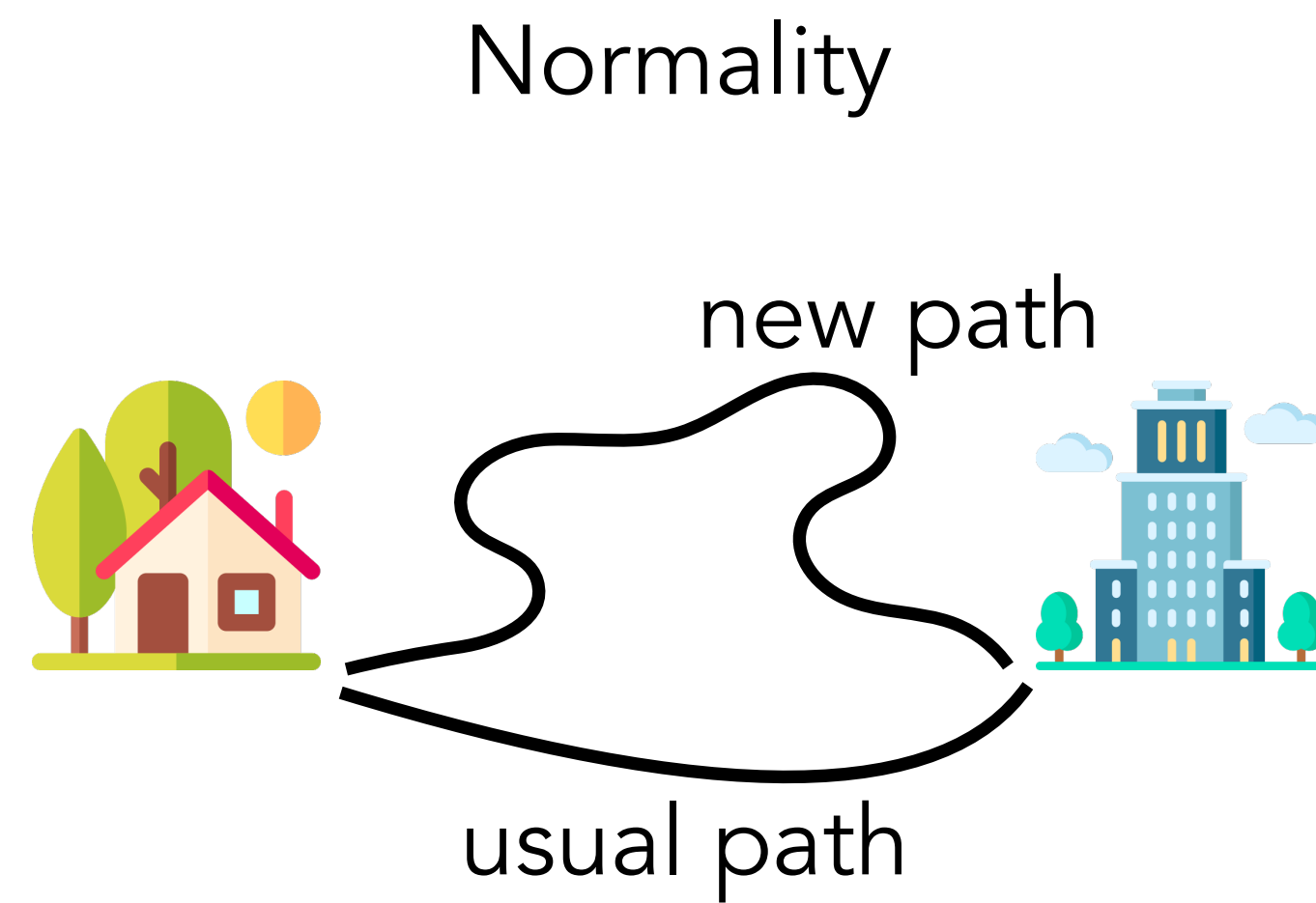
Factors that affect the choice of counterfactual contrasts

Normality



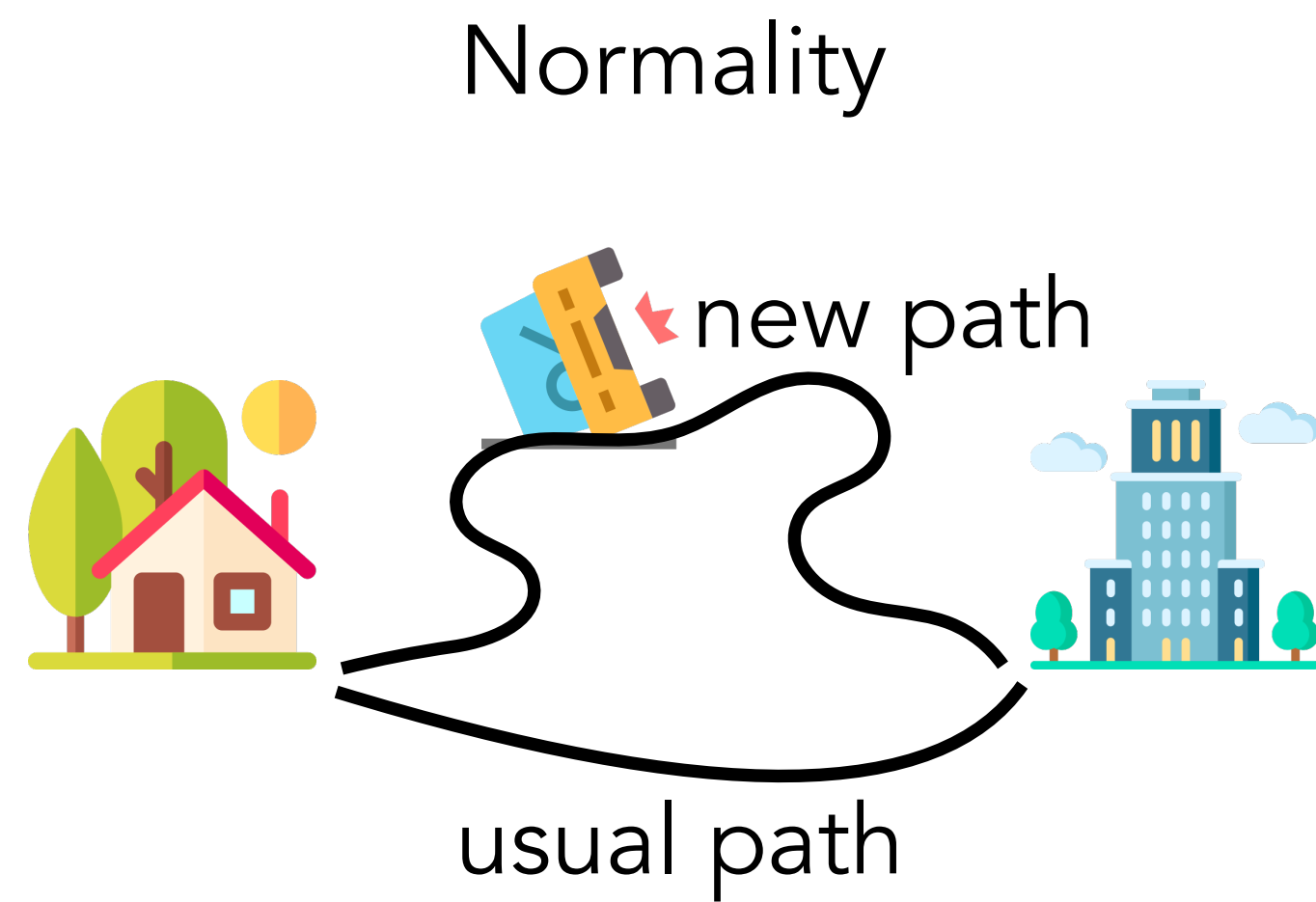
Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" *Psychological review*, 1986.

Factors that affect the choice of counterfactual contrasts



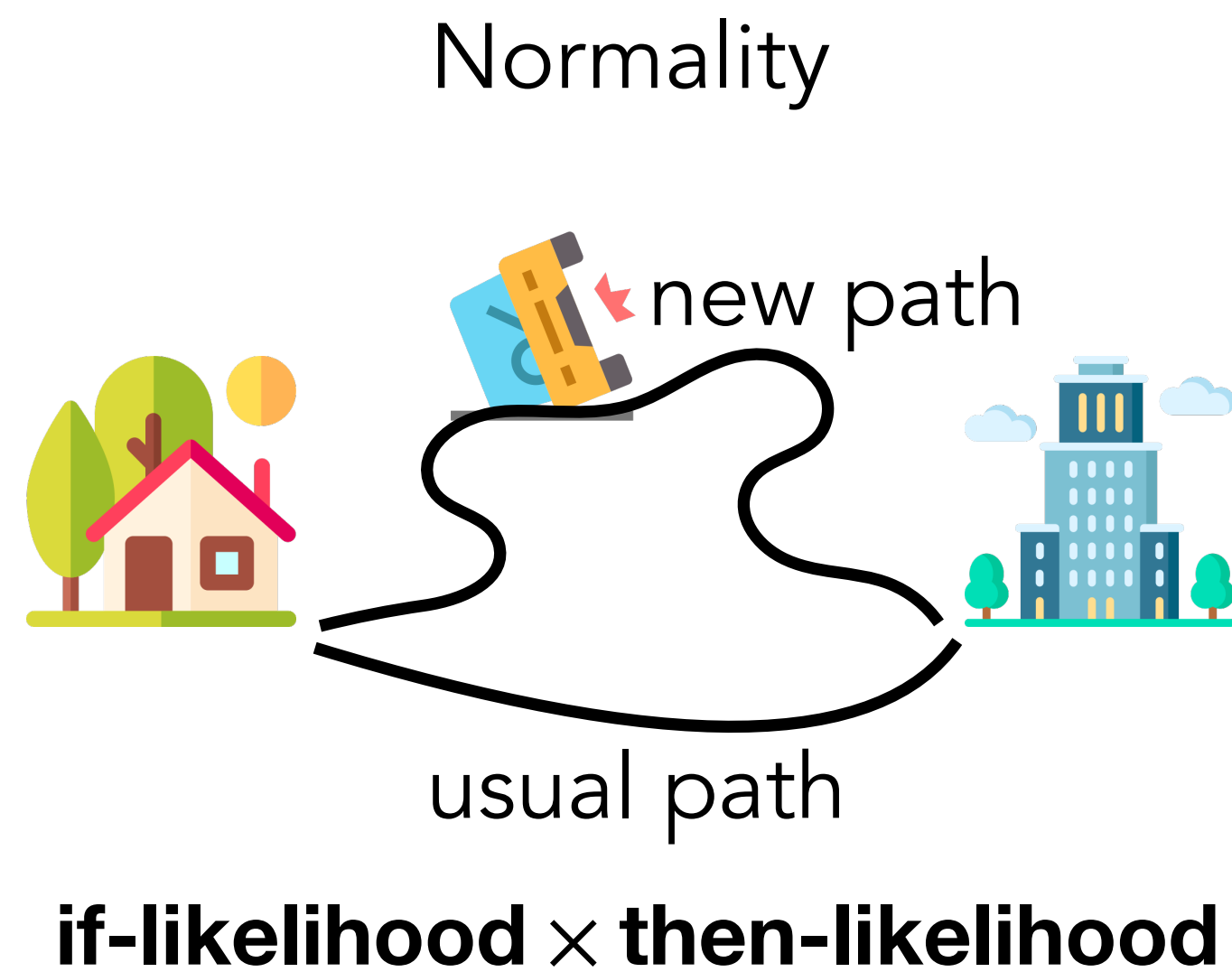
Kahneman and Miller. "*Norm theory: Comparing reality to its alternatives.*" *Psychological review*, 1986.

Factors that affect the choice of counterfactual contrasts



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

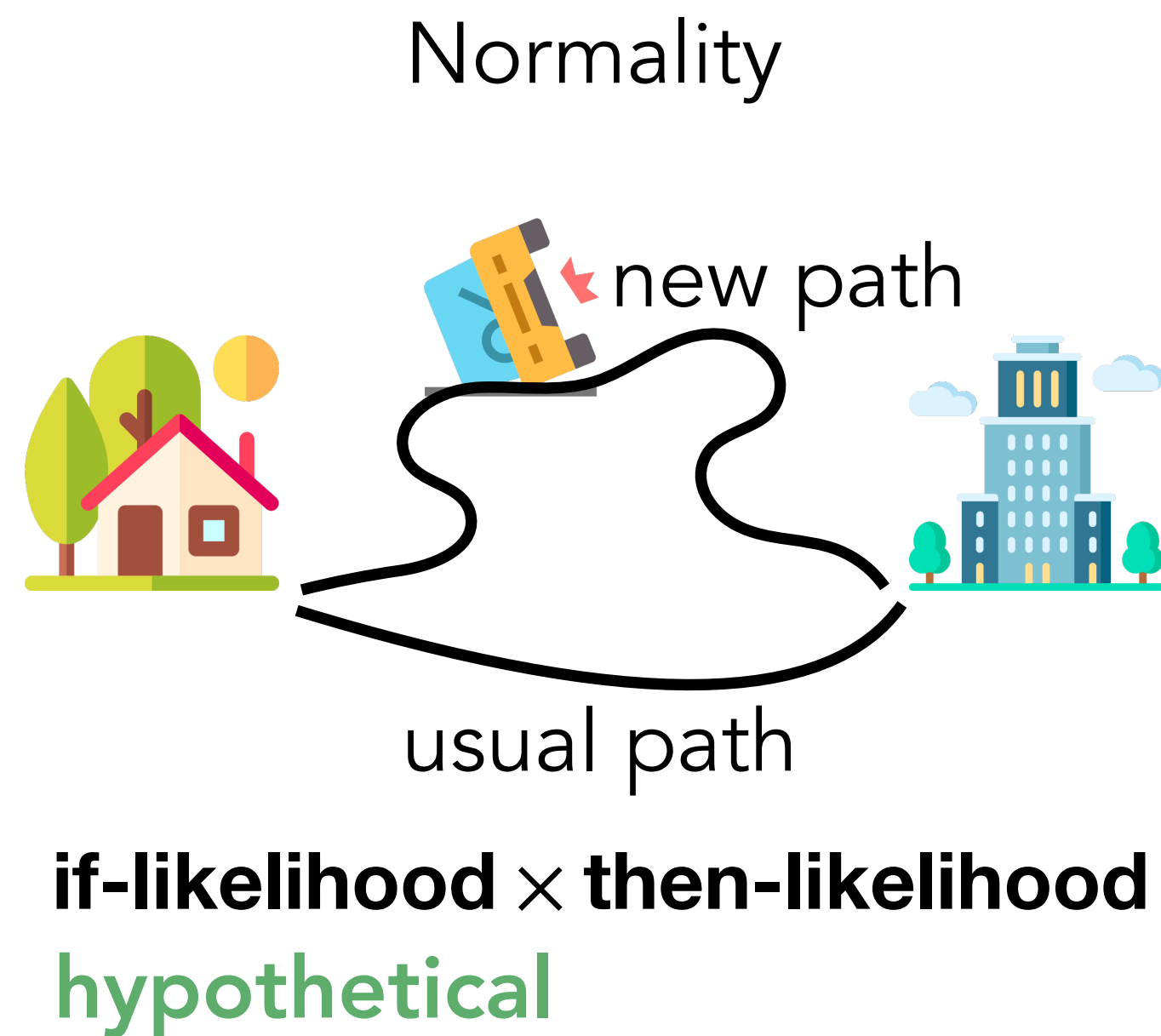
Factors that affect the choice of counterfactual contrasts



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.

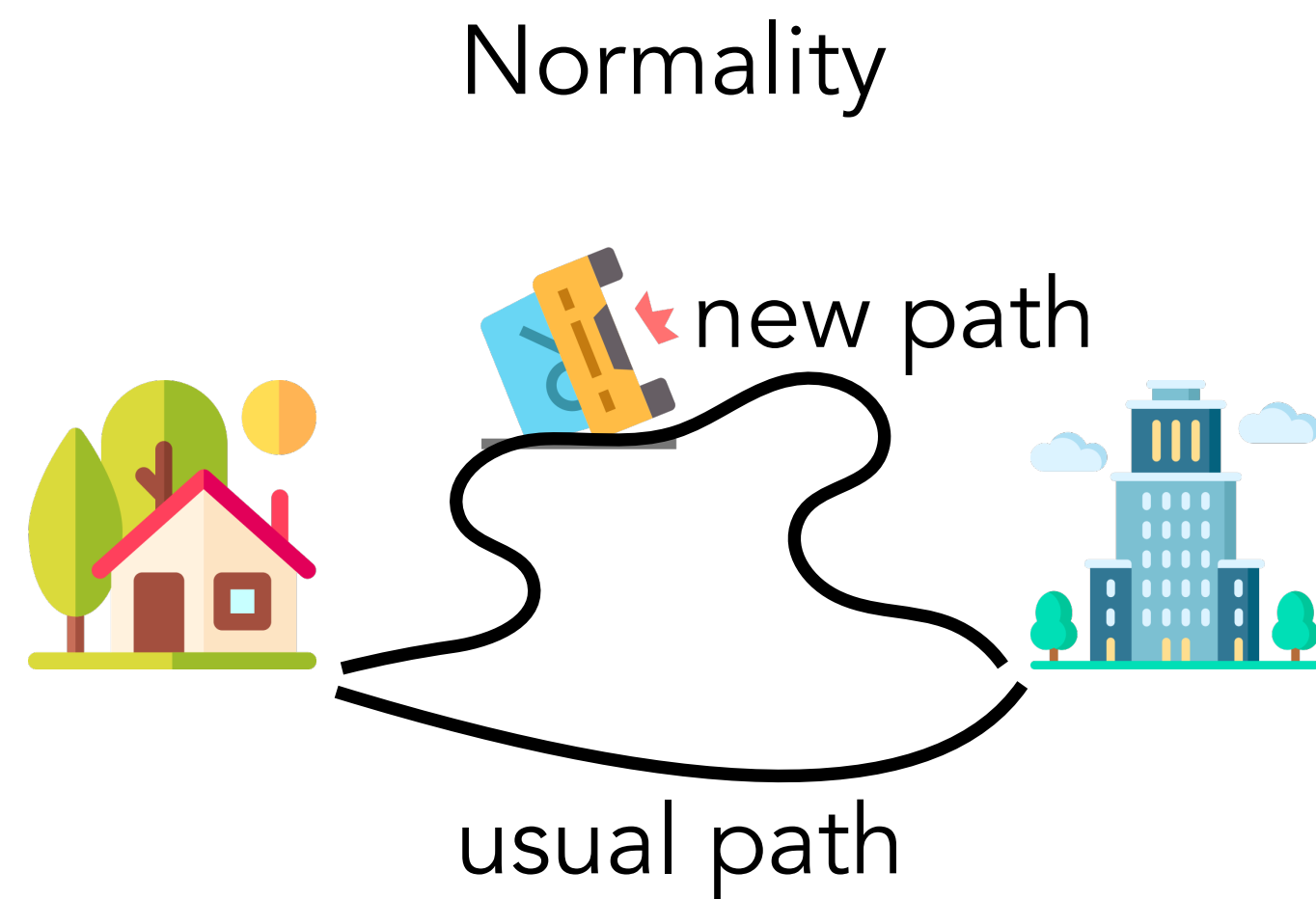
Factors that affect the choice of counterfactual contrasts



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



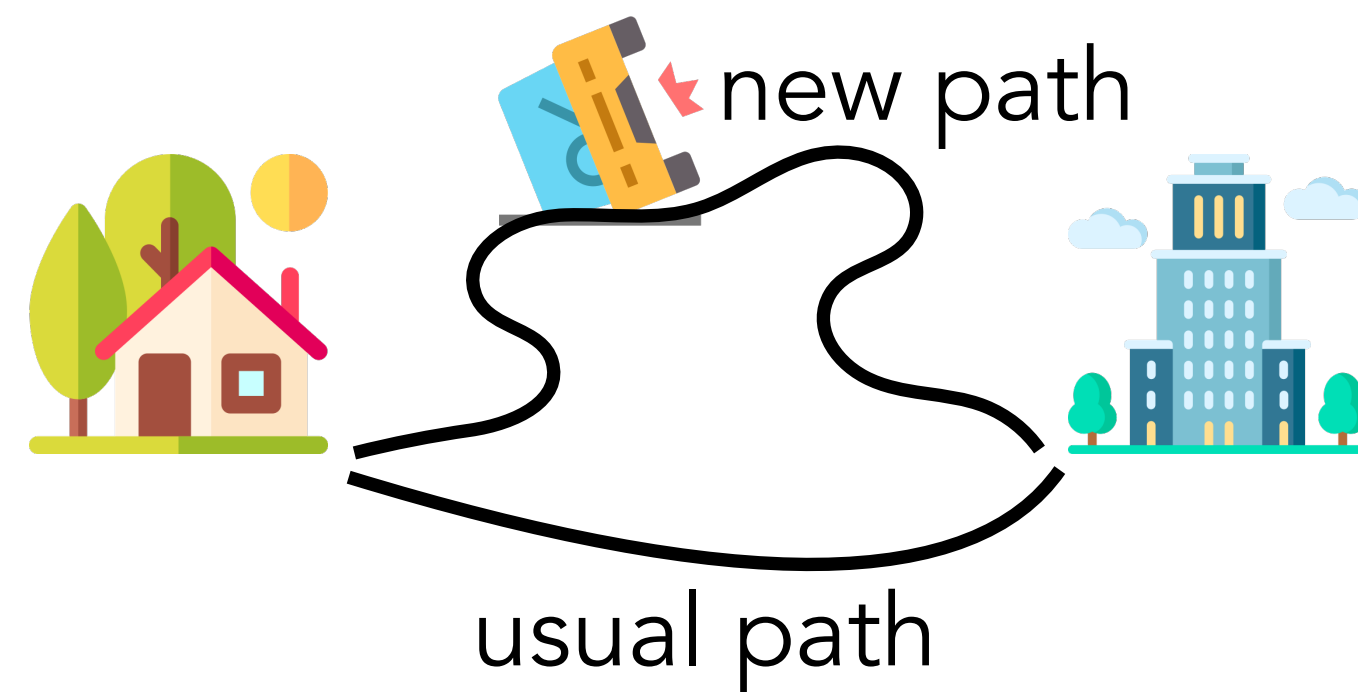
if-likelihood × **then-likelihood**
hypothetical **counterfactual**

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



Recency



if-likelihood × then-likelihood
hypothetical counterfactual

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

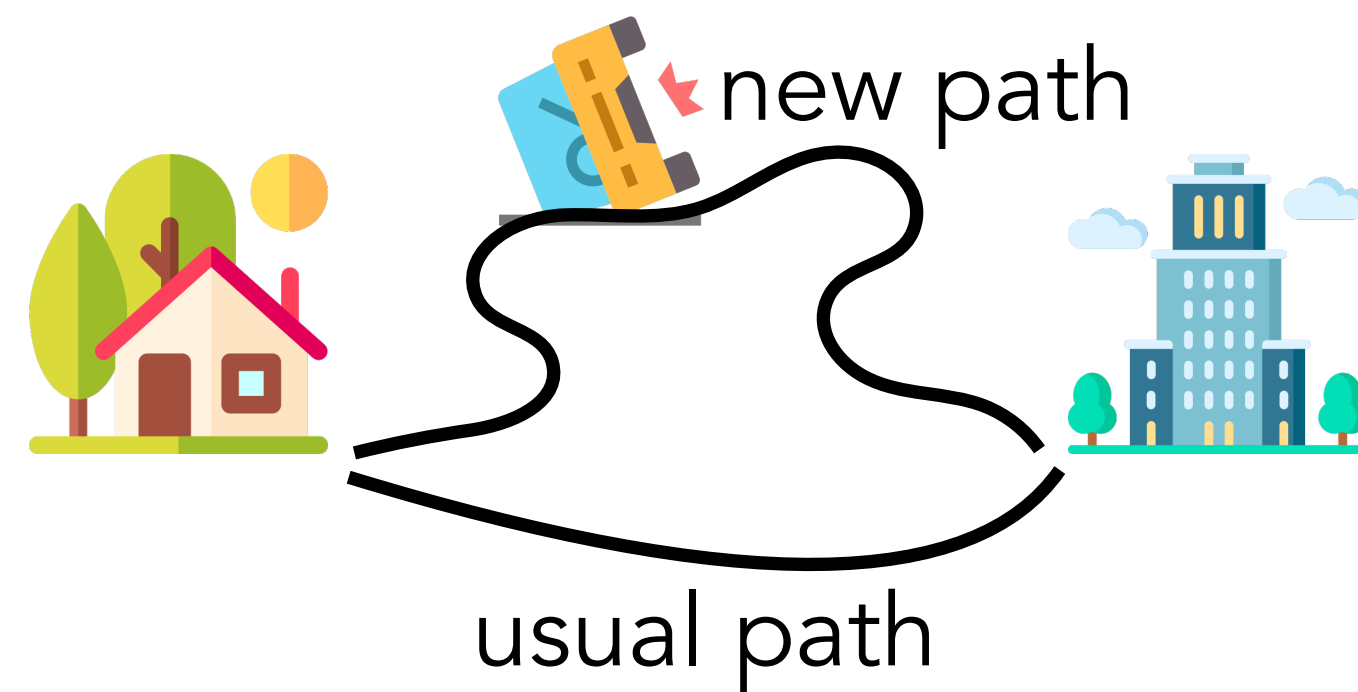
Petrocelli et al. "Counterfactual potency." *Journal of personality and social psychology*, 2011.

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



Recency



Controllability

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.

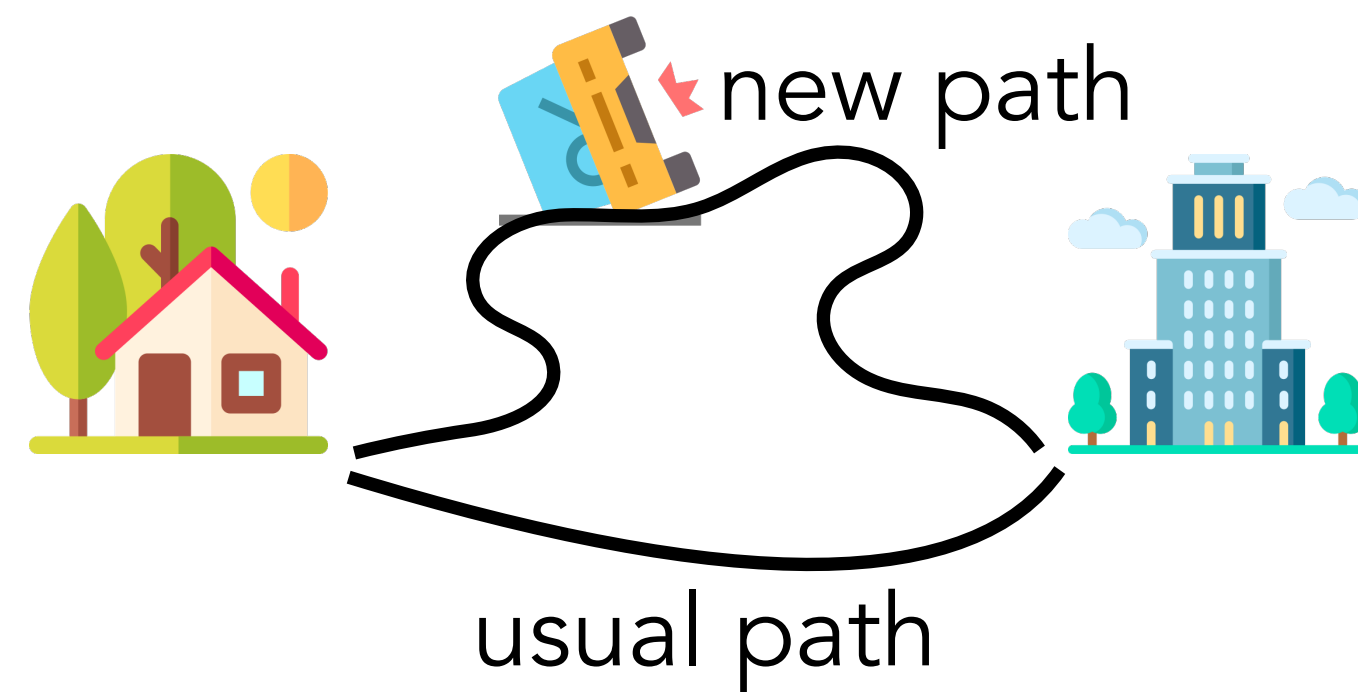
Spellman. "Crediting causality." Journal of Experimental Psychology, 1997.

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

Giroto 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.

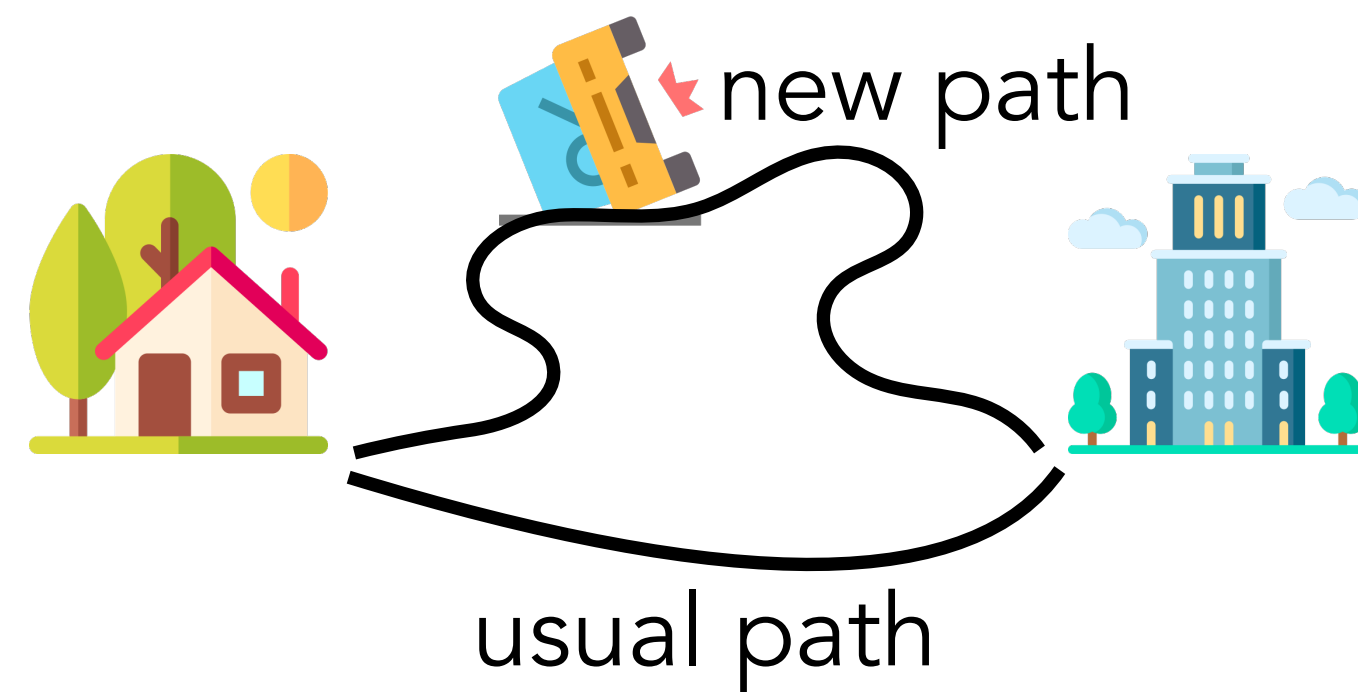
Spellman. "Crediting causality." Journal of Experimental Psychology, 1997.

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

Giroto 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.

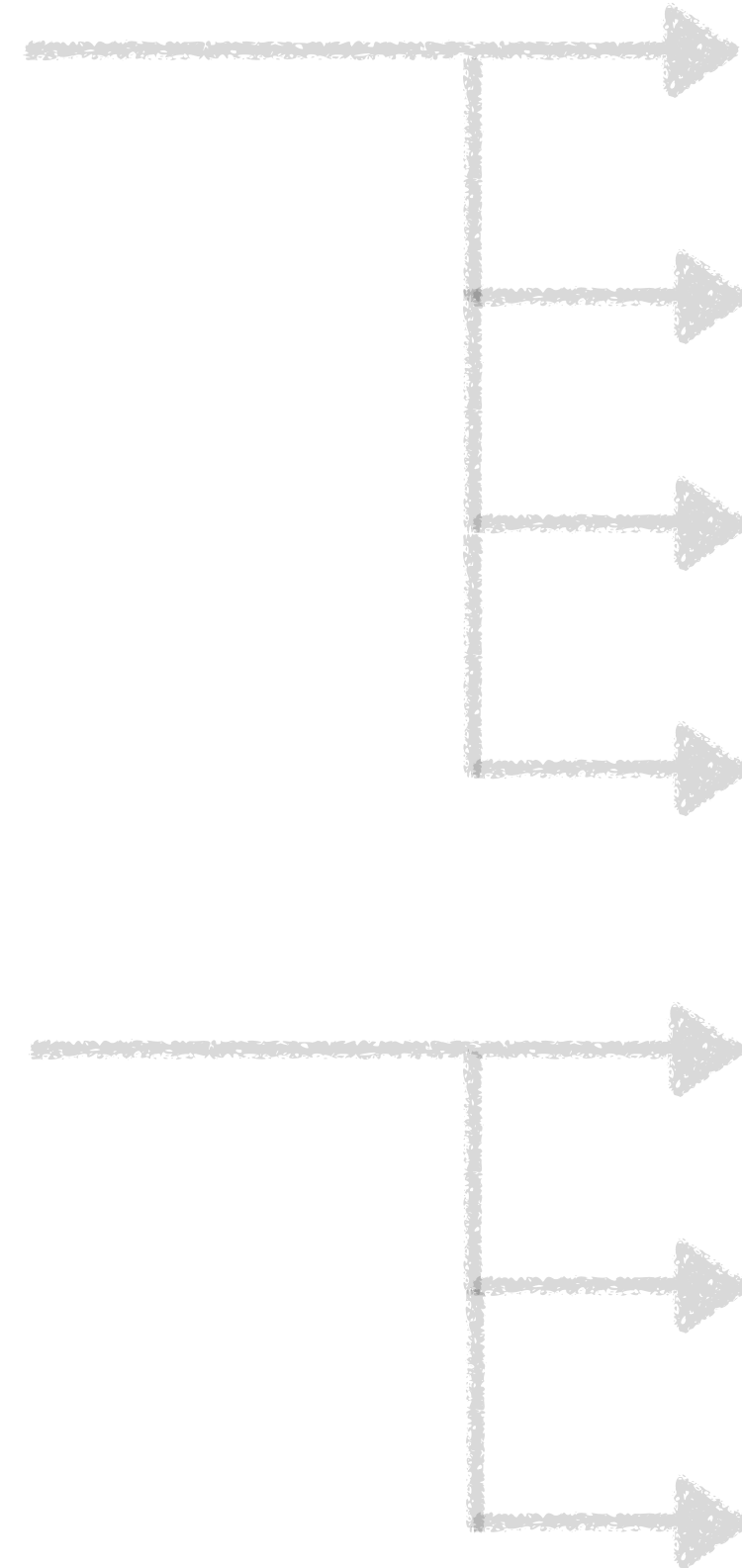
Petrocelli et al. "Counterfactual potency." Journal of personality and social psychology, 2011.

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

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

We think of counterfactuals all the time



How? → *Mental simulation*

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.

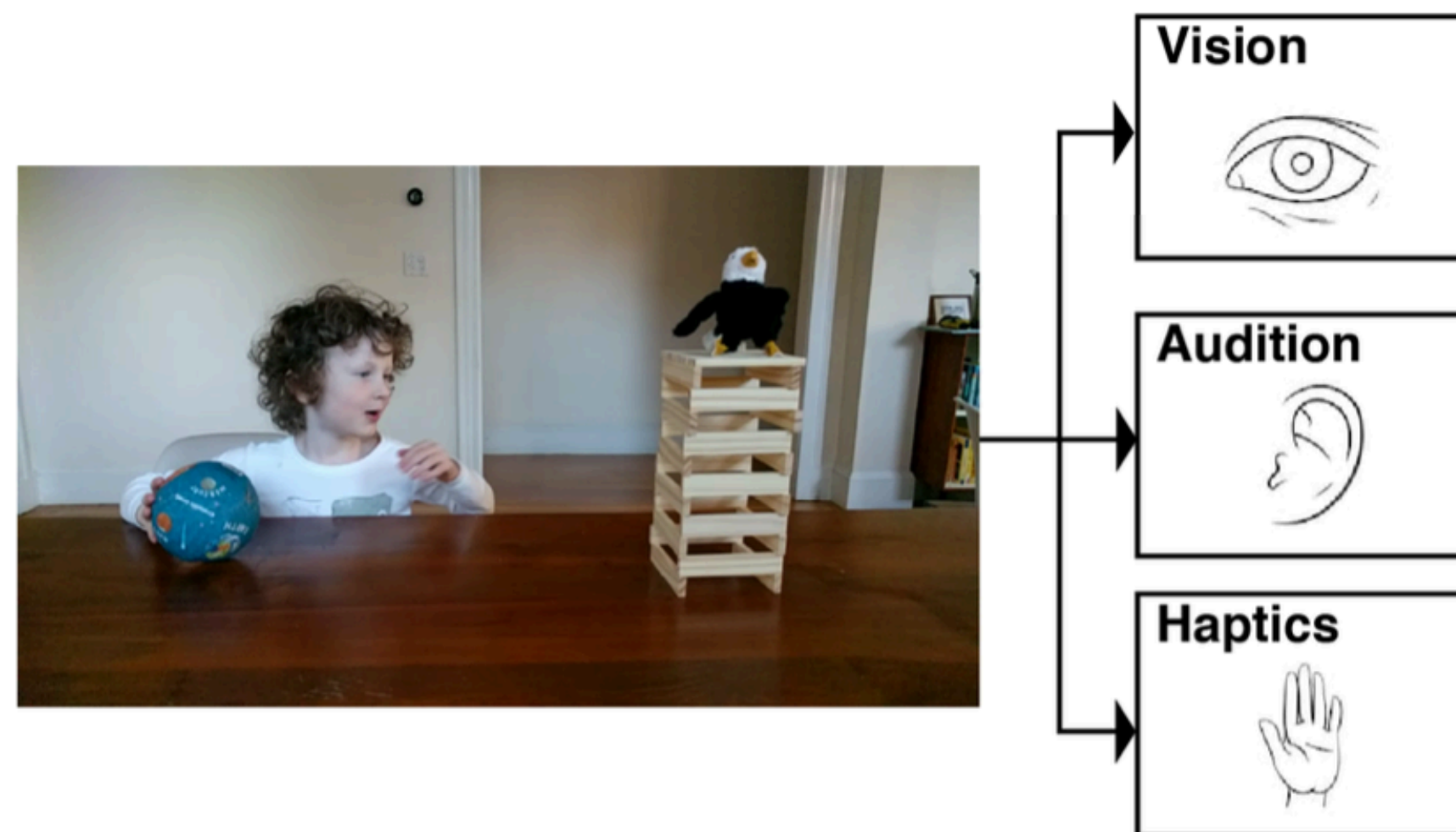
To advance the study of availability for construction, we now sketch a mental operation that we label the simulation heuristic. Our starting point is a common introspection: **There appear to be many situations in which questions about events are answered by an operation that resembles the running of 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.

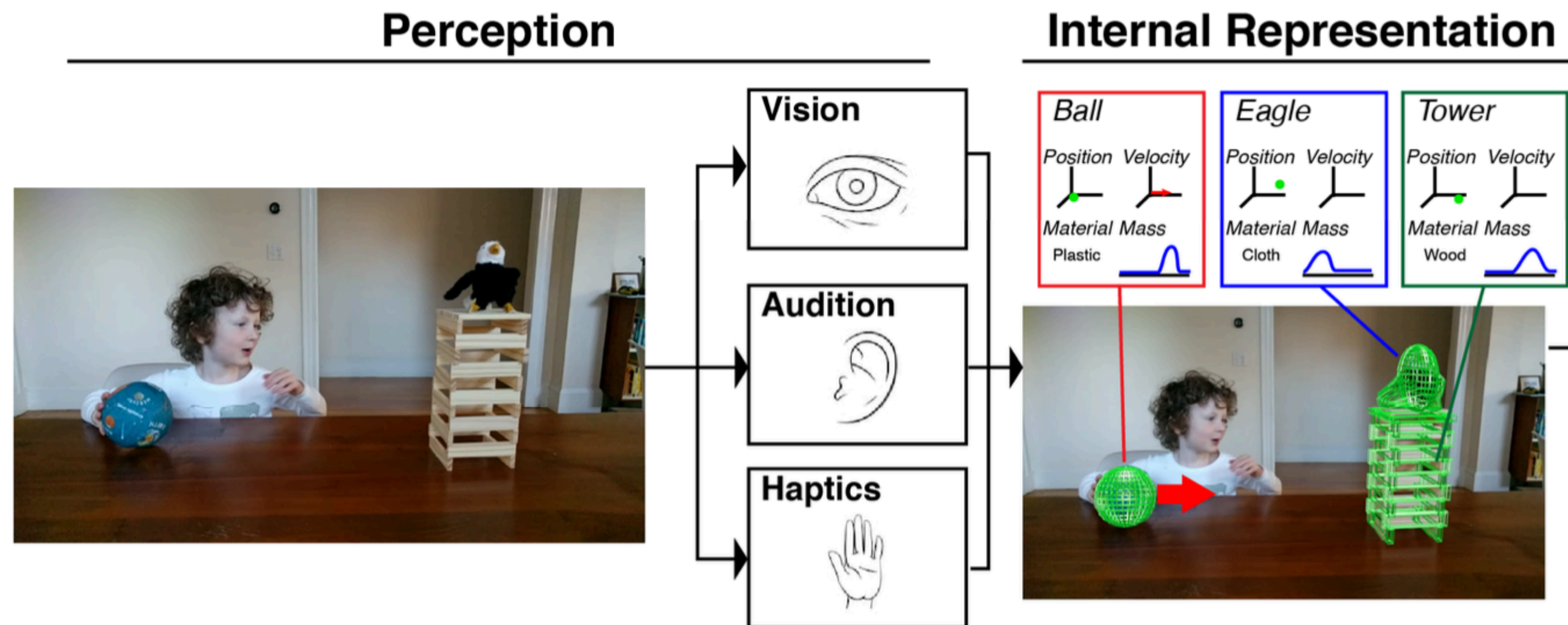
Mental machinery and operations

Perception



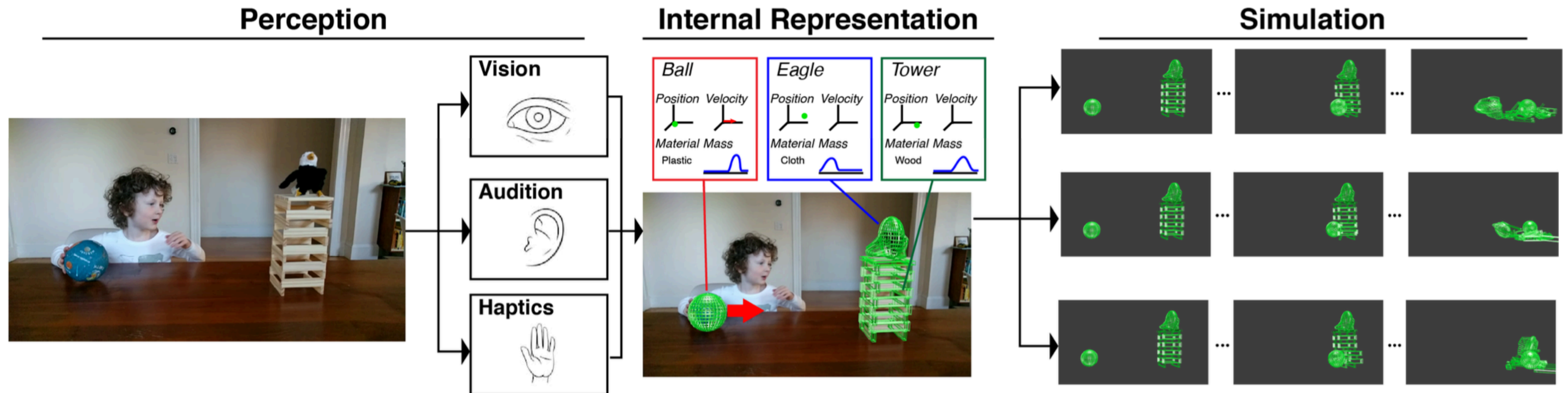
Smith et al. "*Probabilistic models of physical reasoning.*" In Bayesian Models of Cognition: Reverse Engineering the Mind, MIT Press, 2025.

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

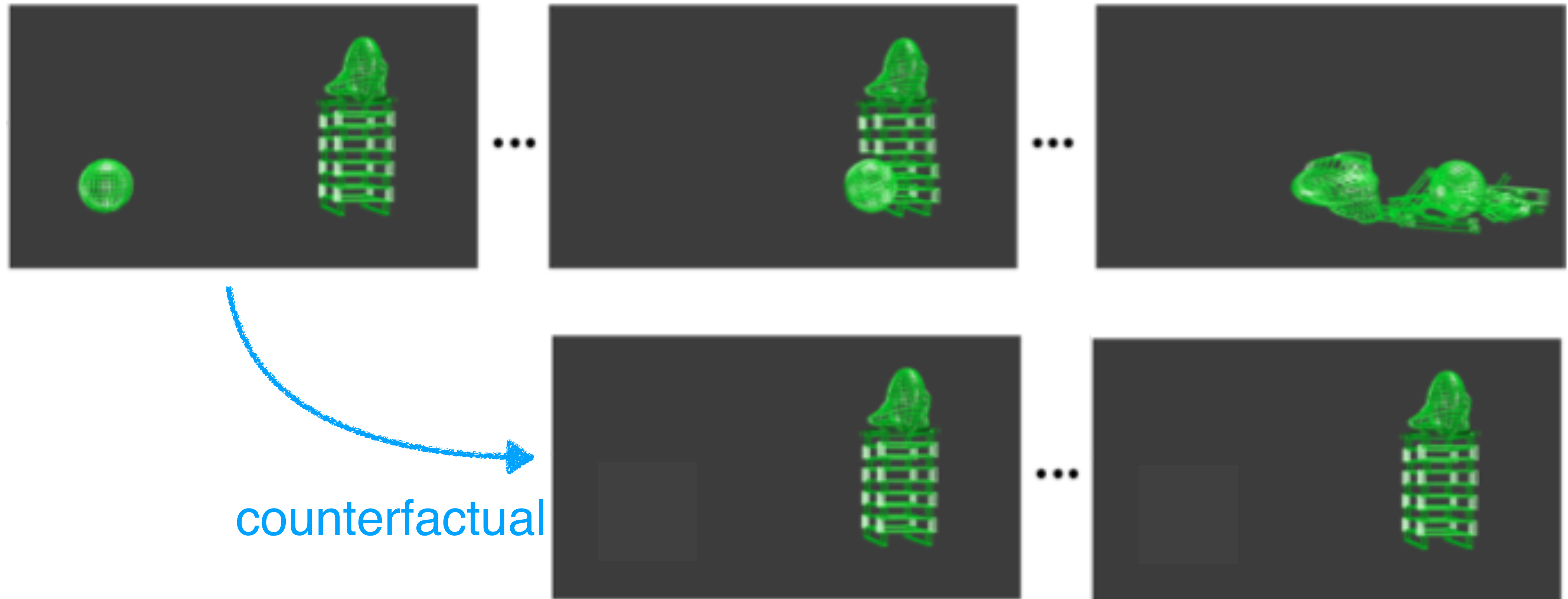
Explain why something happened



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.



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-

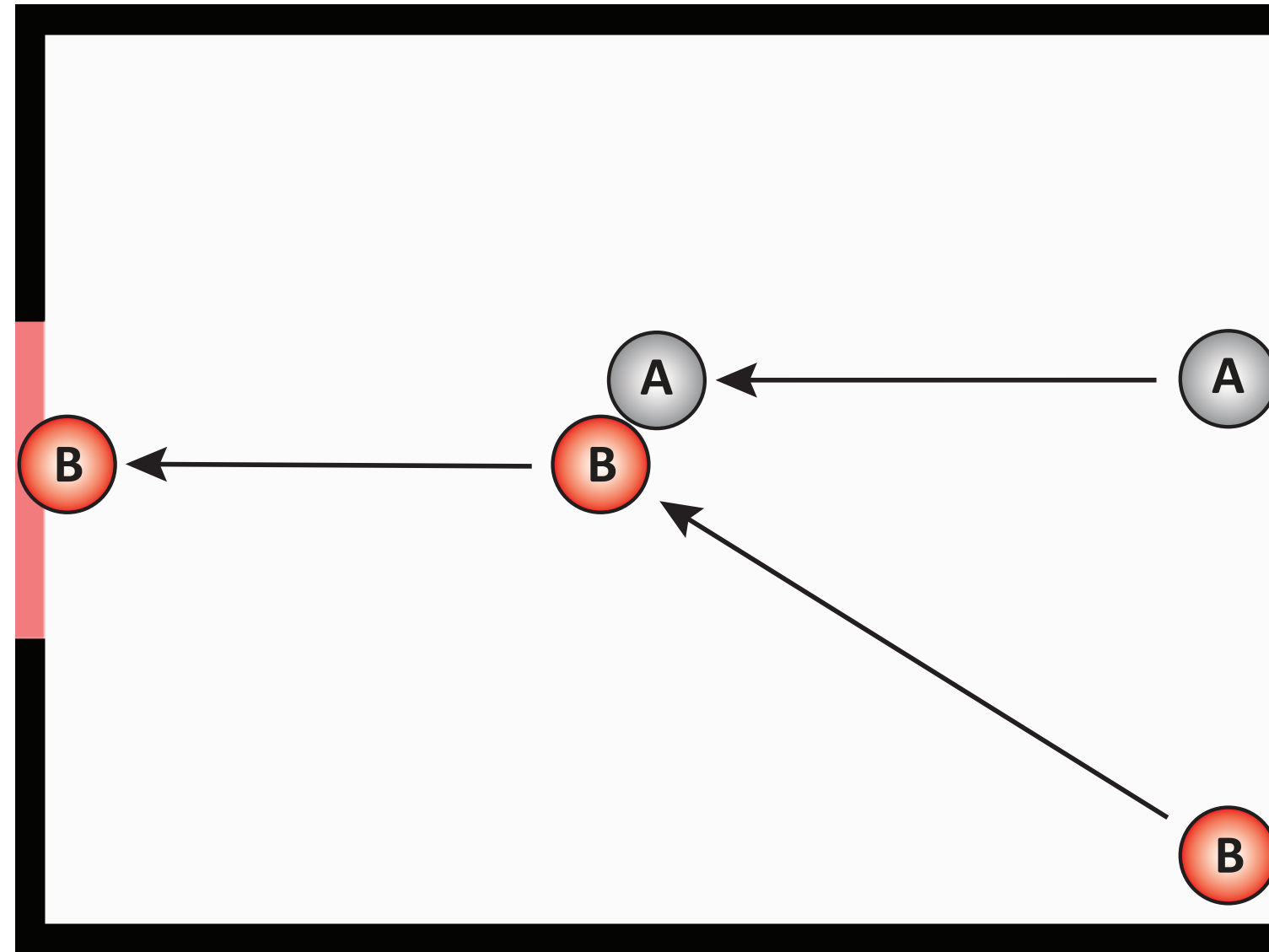
Deep dive: Counterfactual simulation for causal judgments

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

Watch Clip 1

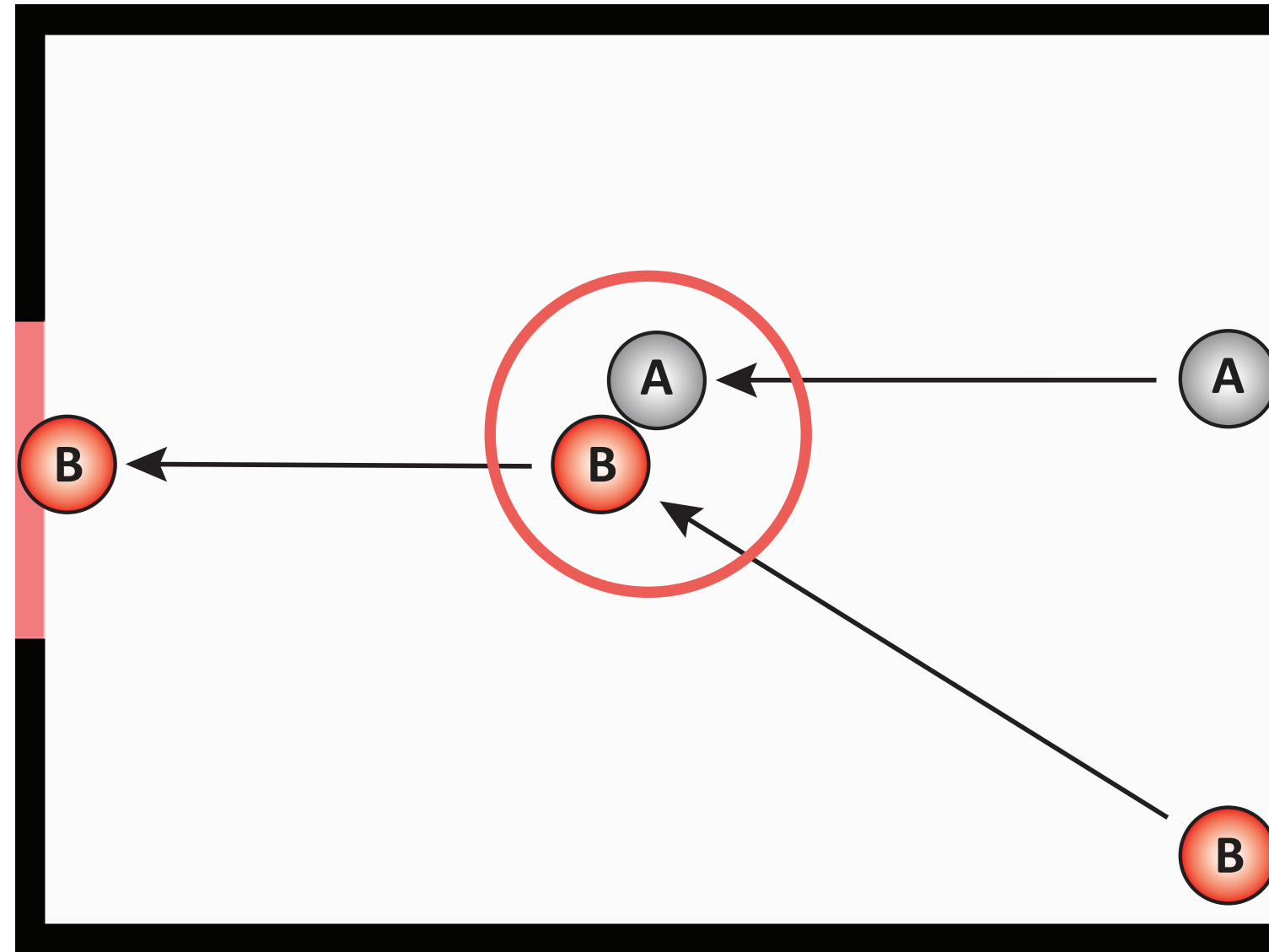
Counterfactual Simulation Model

What happened?



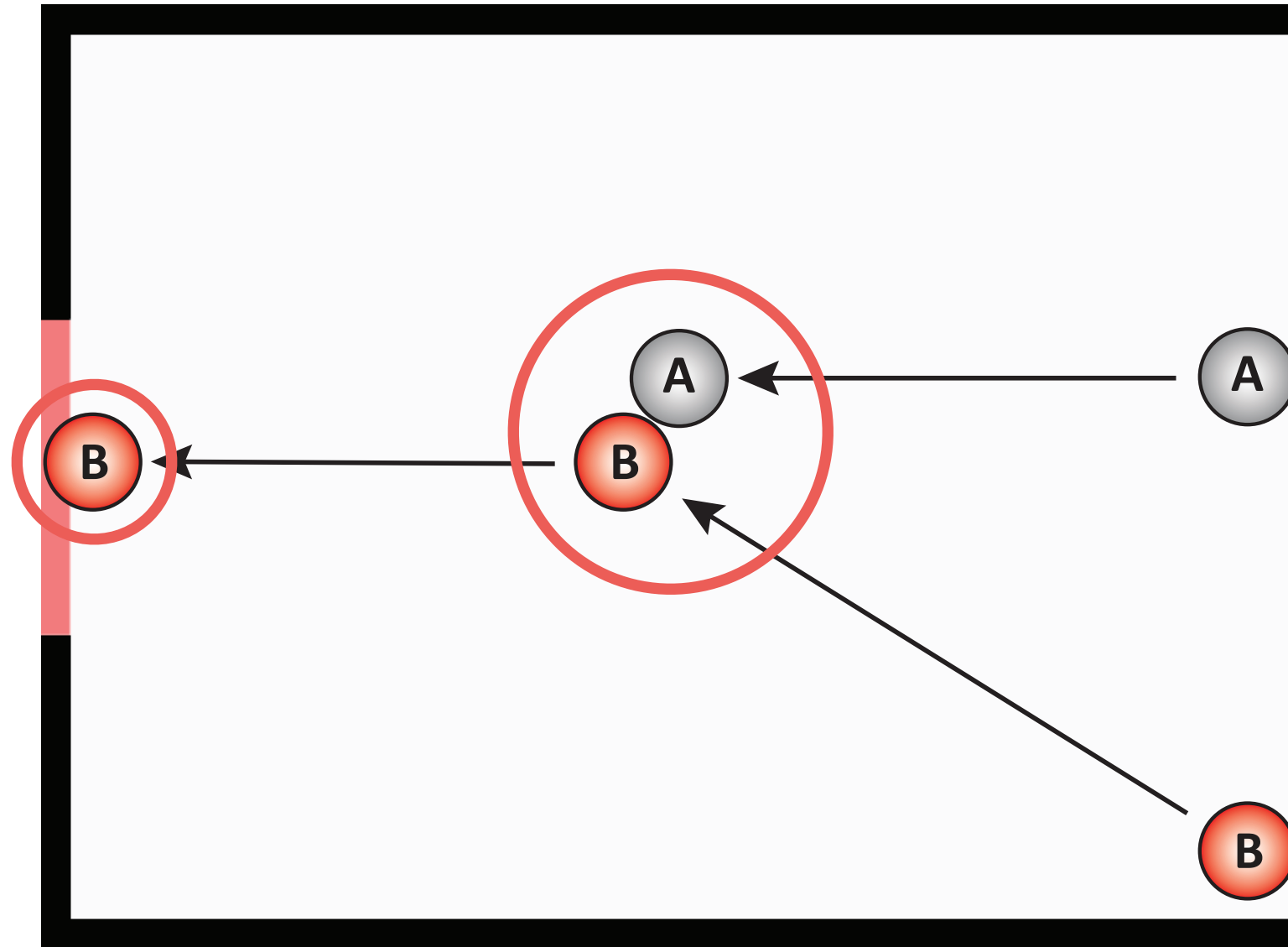
Counterfactual Simulation Model

What happened?



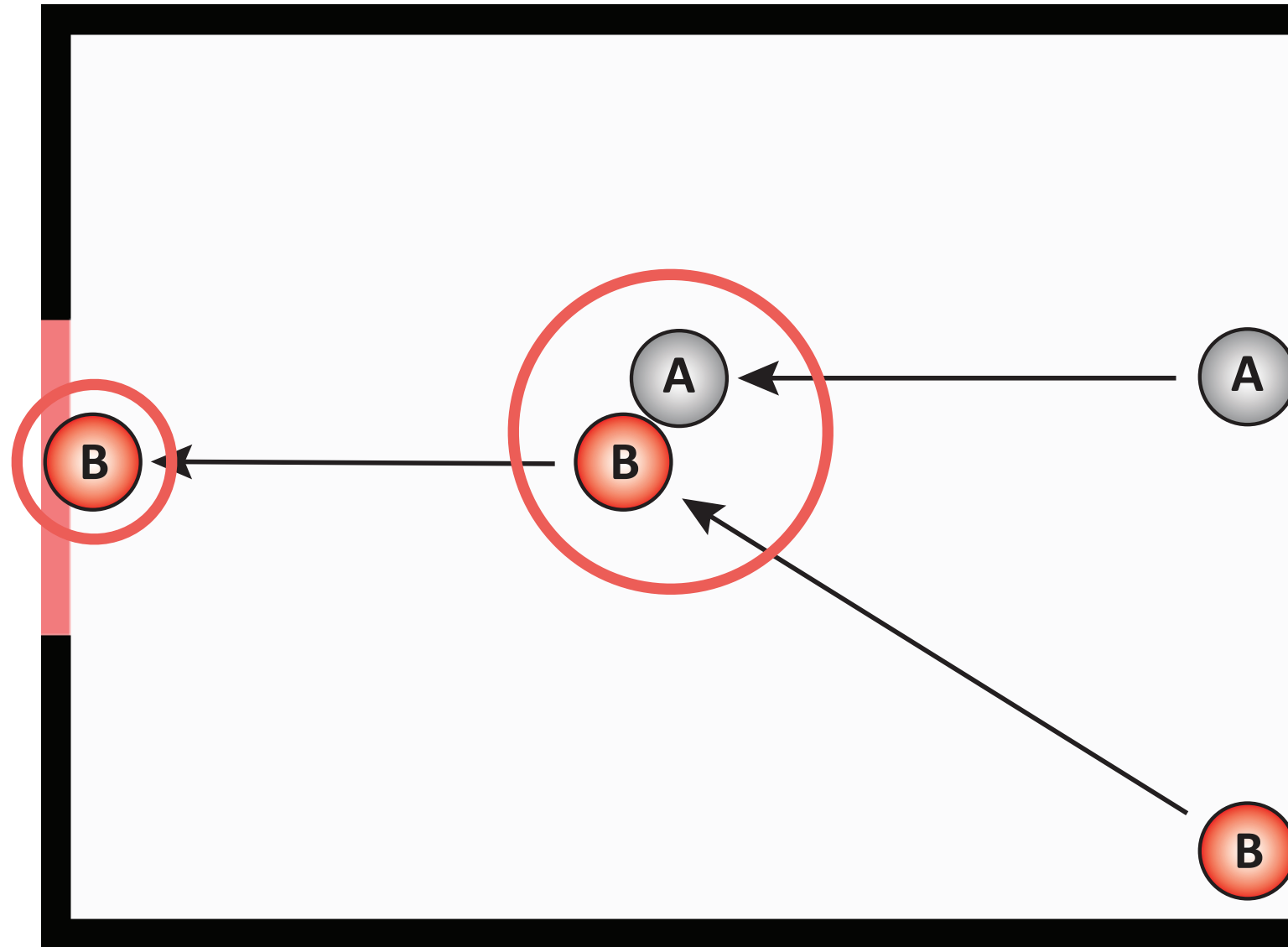
Counterfactual Simulation Model

What happened?

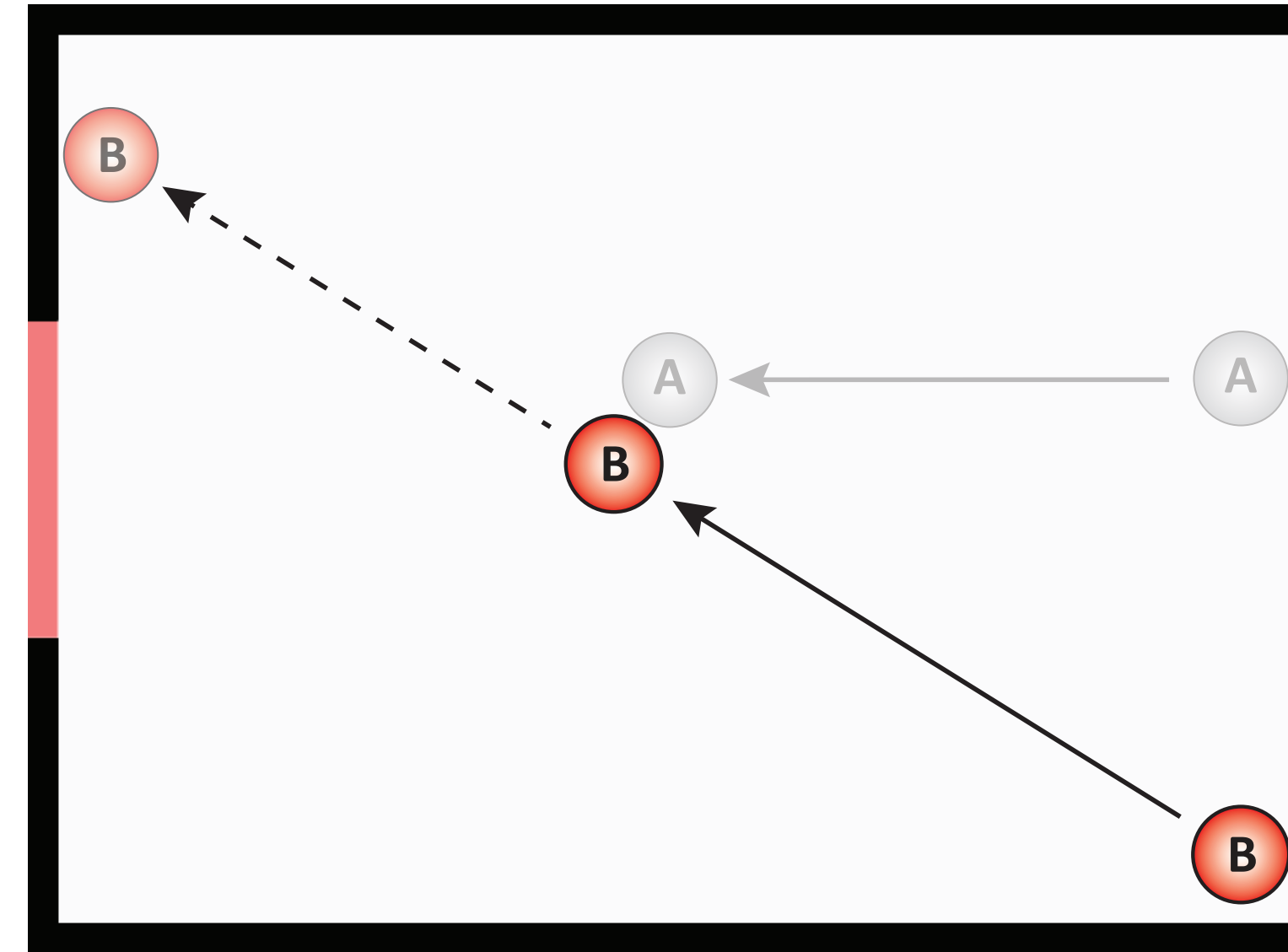


Counterfactual Simulation Model

What happened?

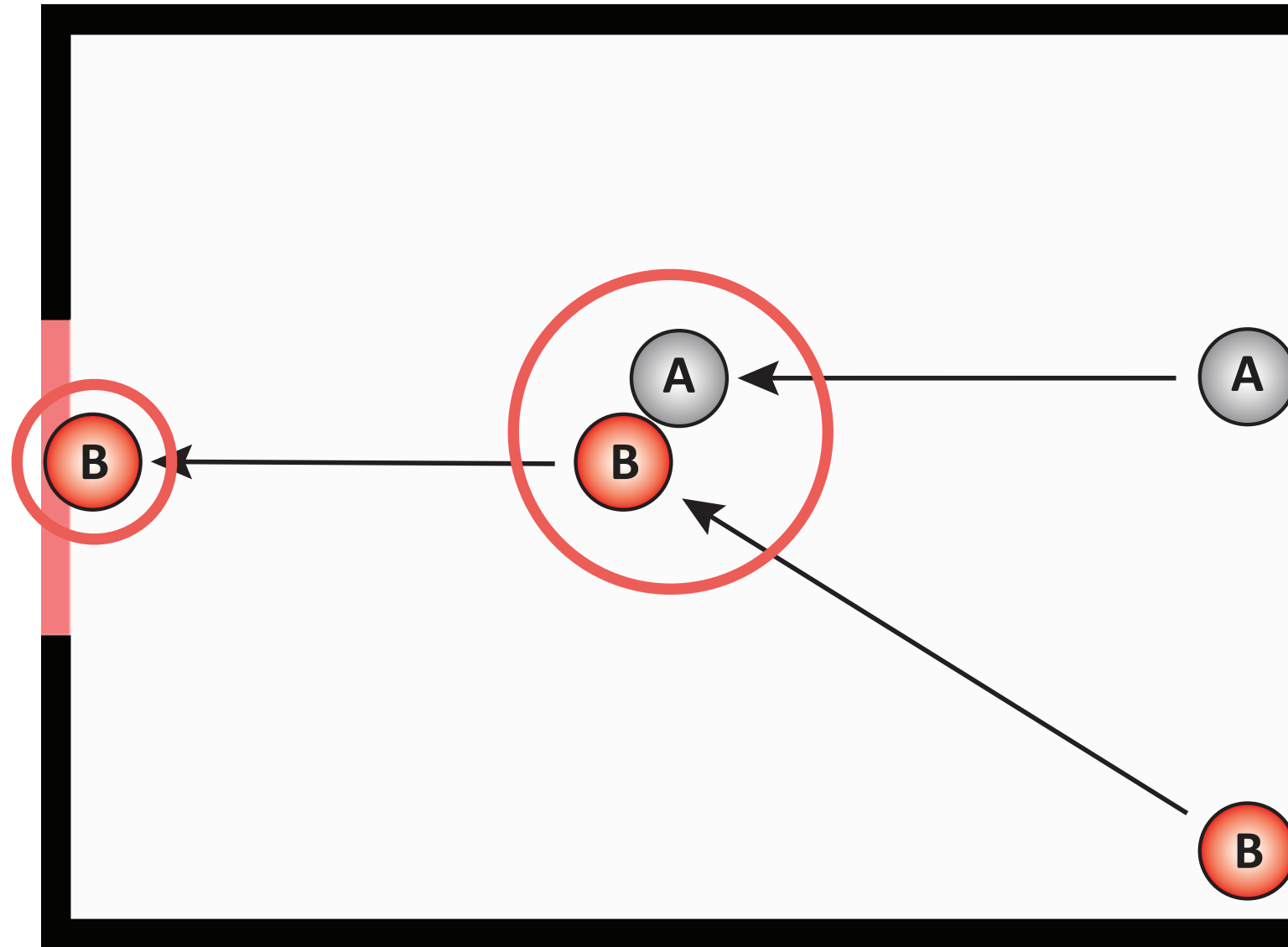


What would have happened?

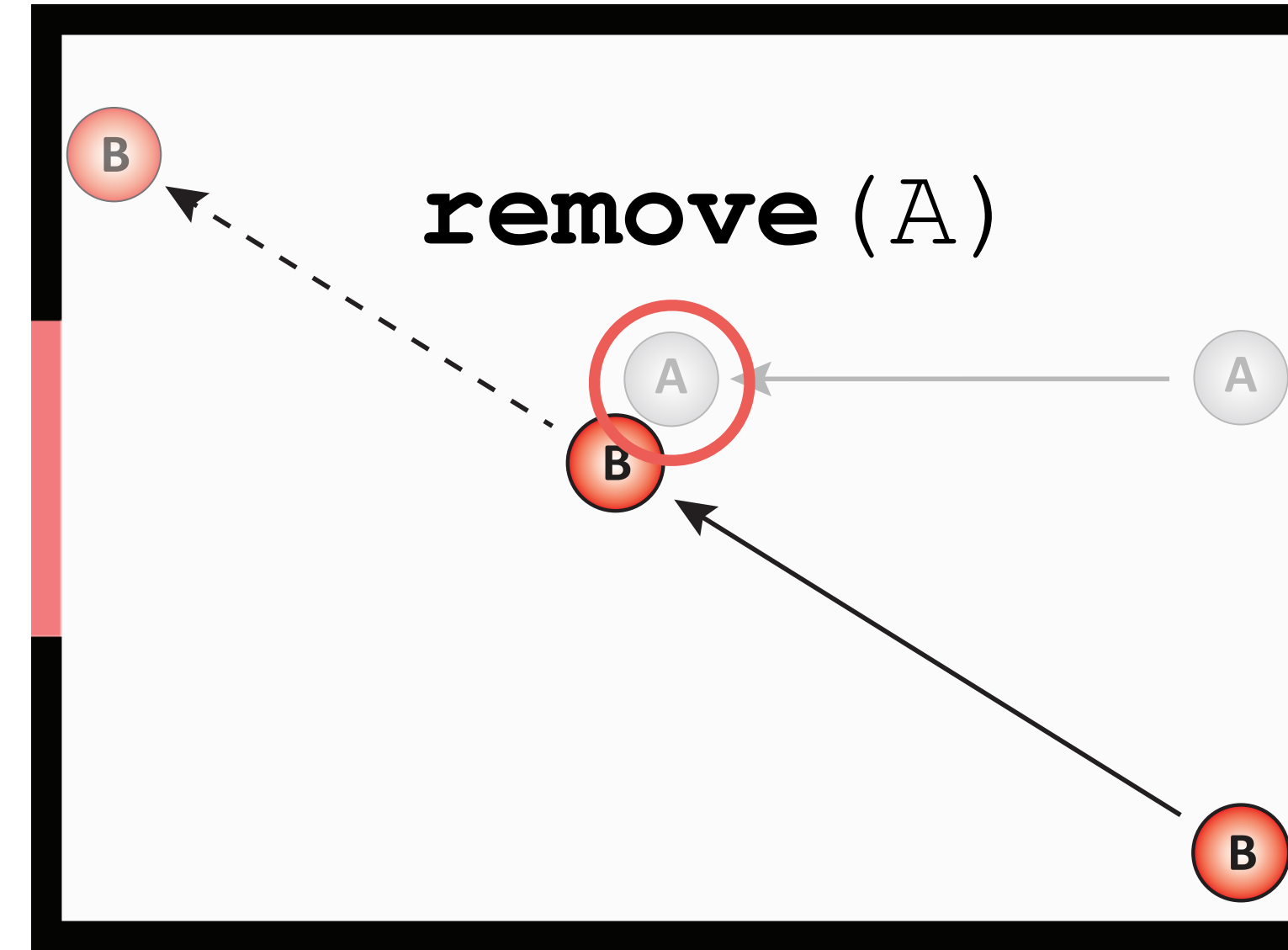


Counterfactual Simulation Model

What happened?

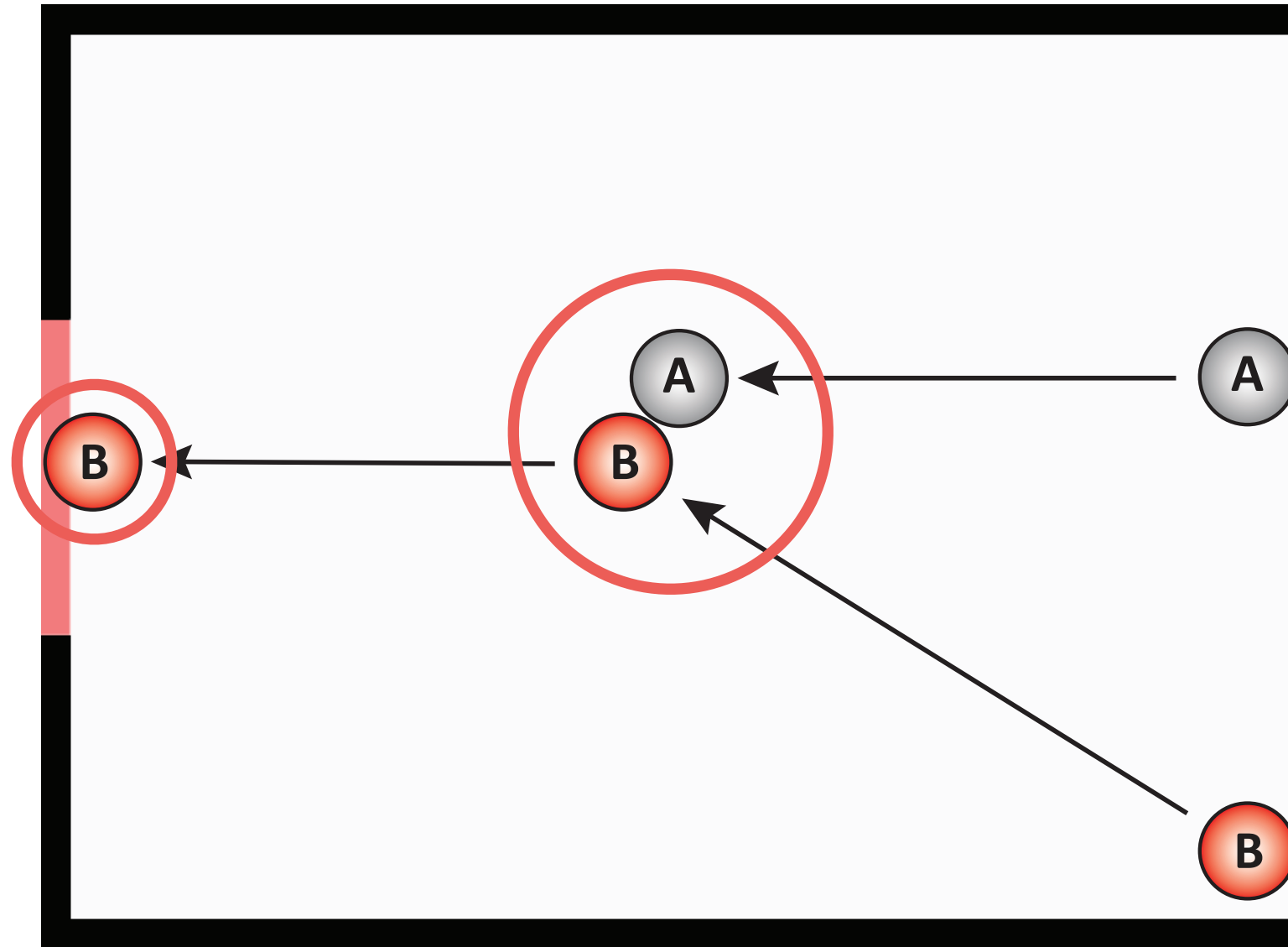


What would have happened?

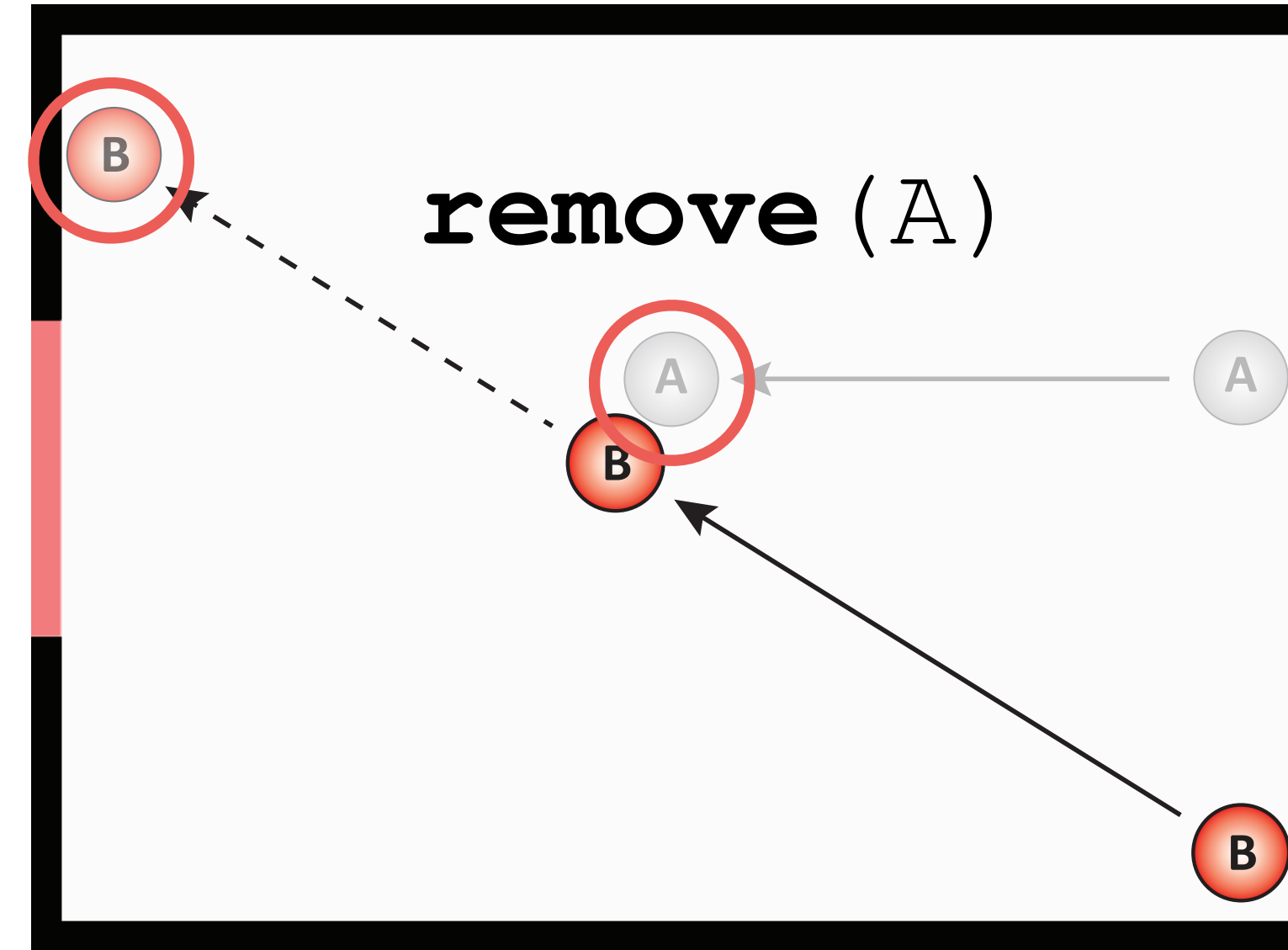


Counterfactual Simulation Model

What happened?

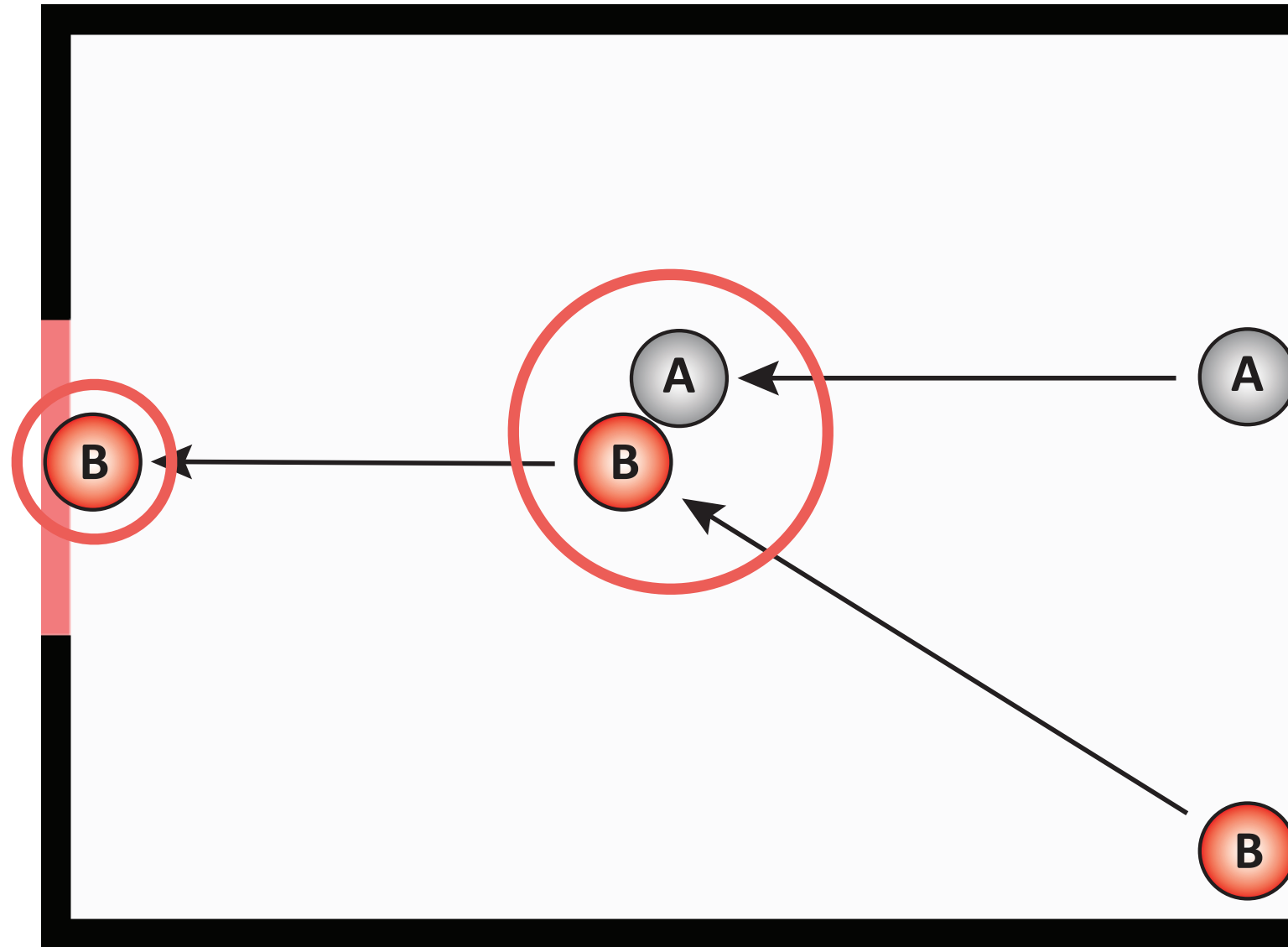


What would have happened?



Counterfactual Simulation Model

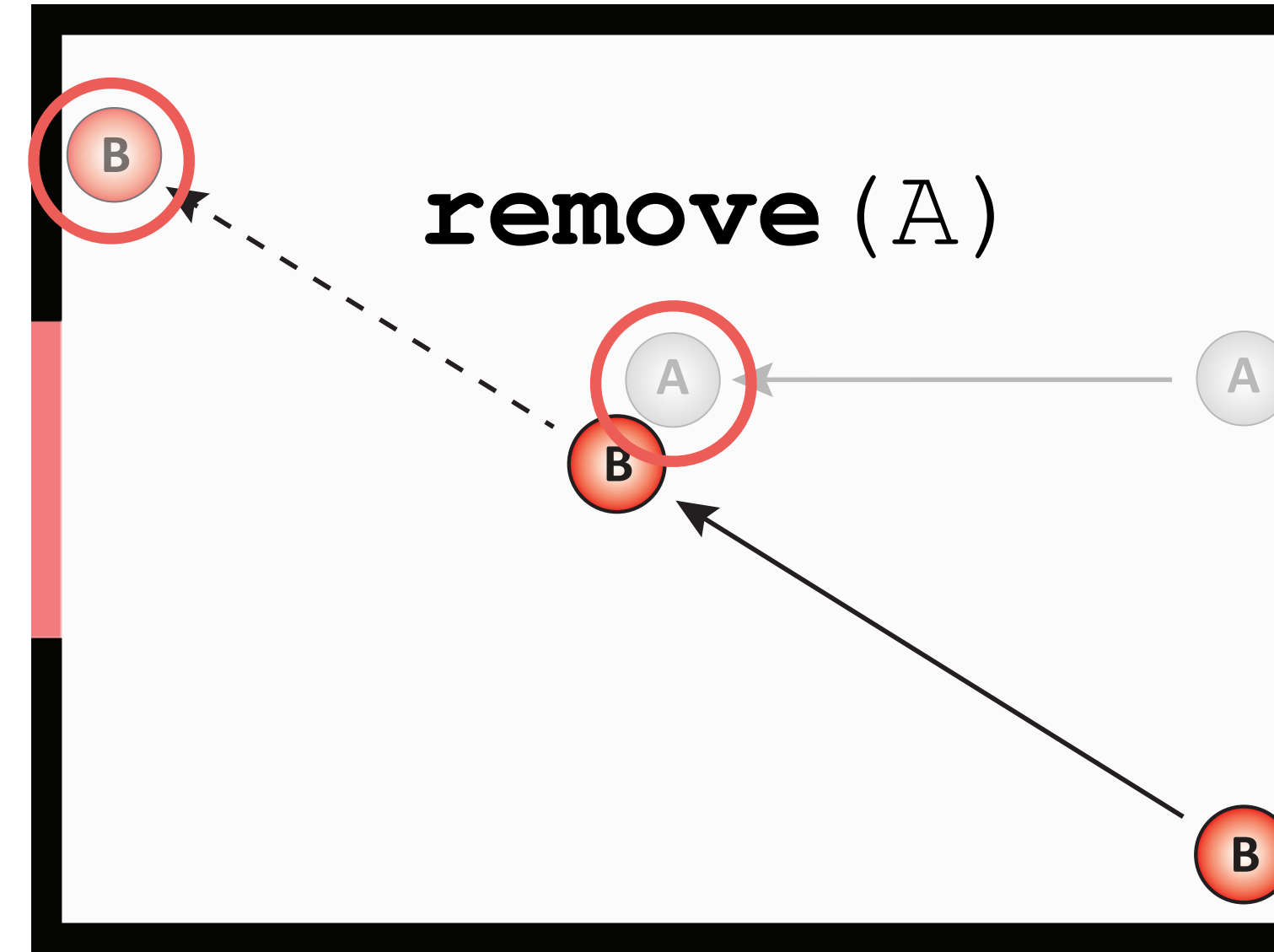
What happened?



Actual situation

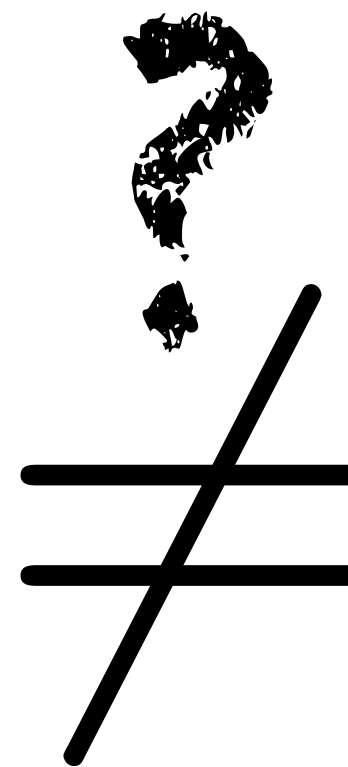
 went through the gate

What would have happened?



Counterfactual situation

 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)
}
```

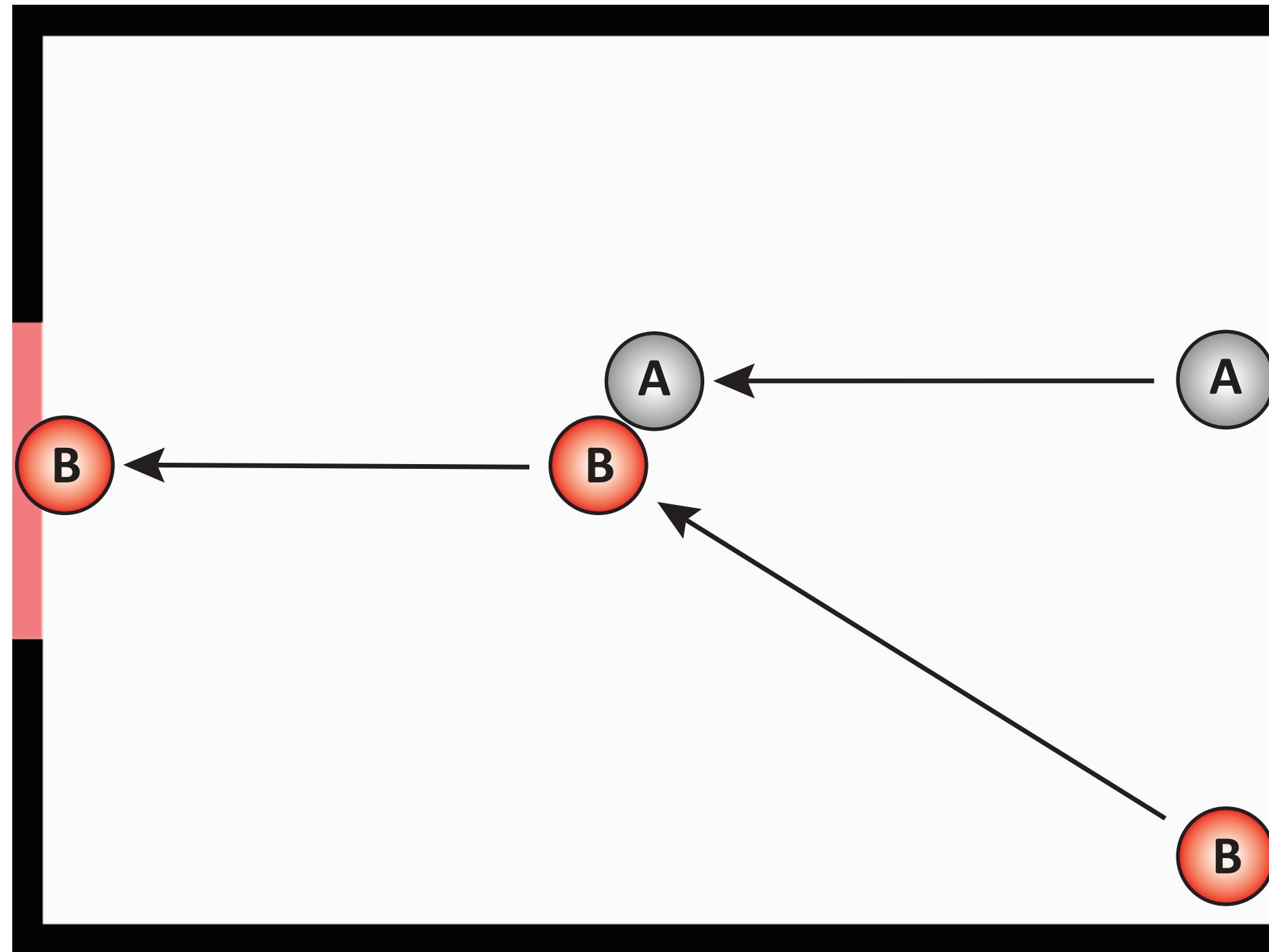
Counterfactual intervention

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.

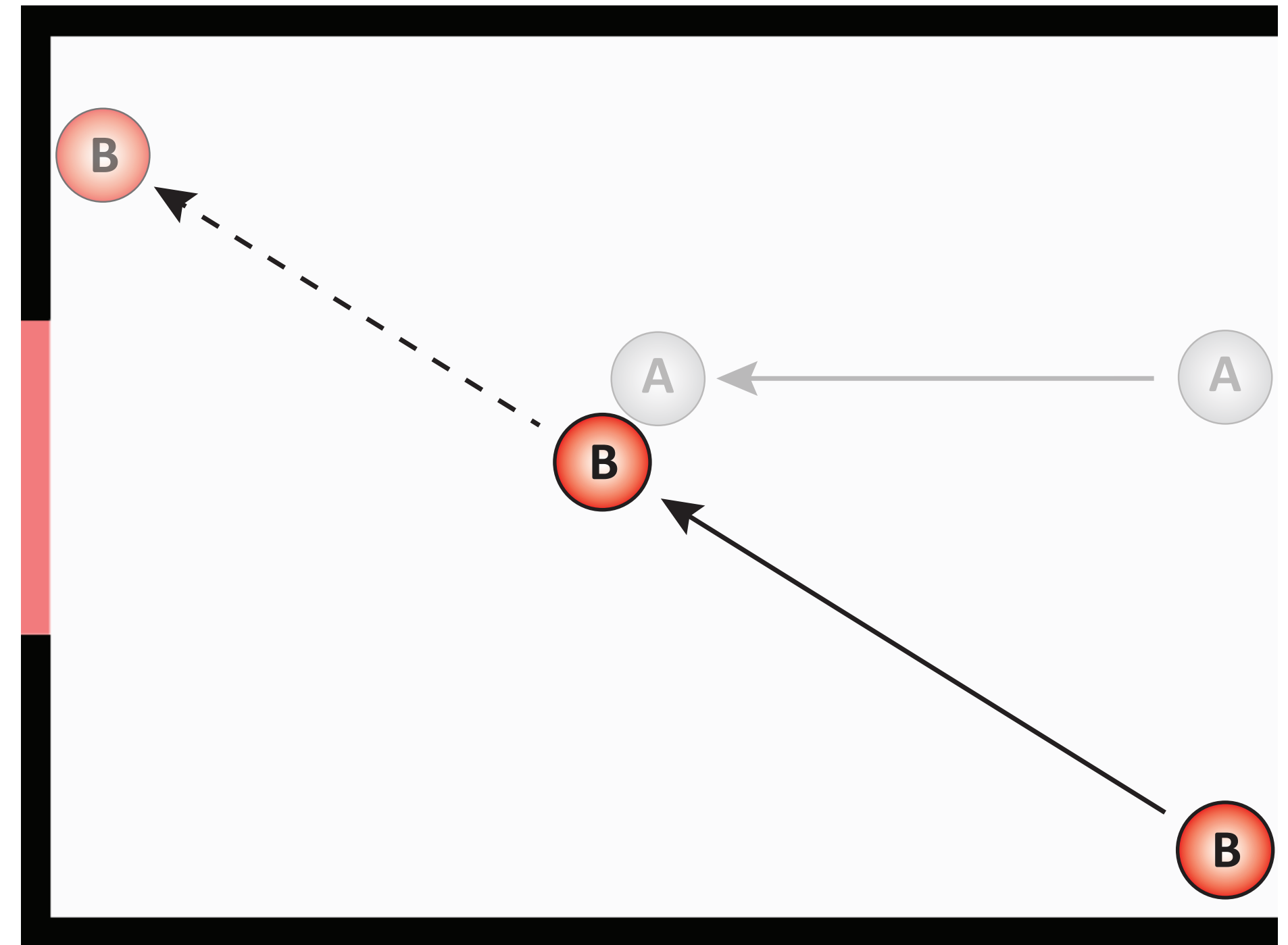
What happened?



Actual situation

 went through the gate

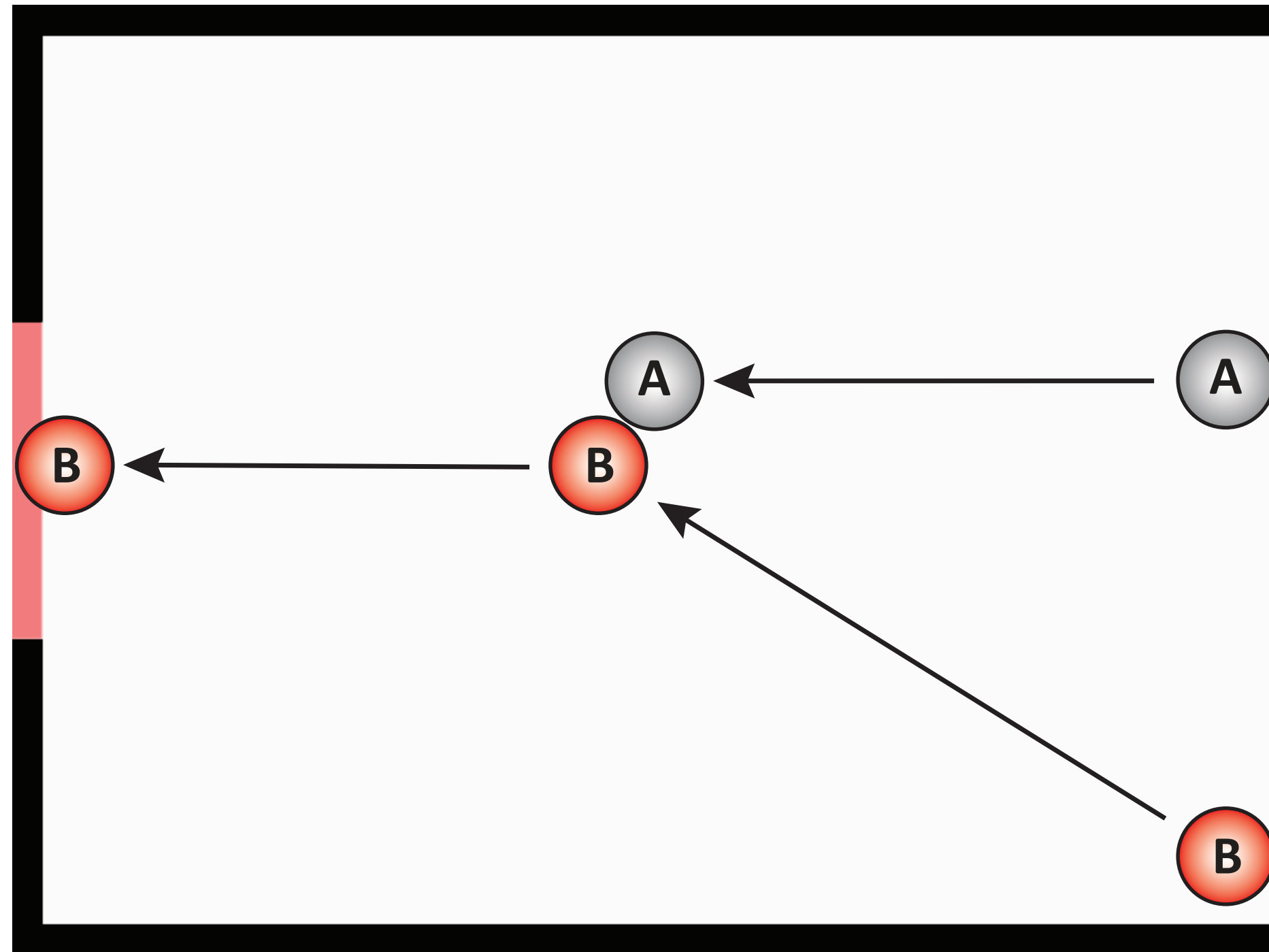
What would have happened?



Counterfactual situation

\neq

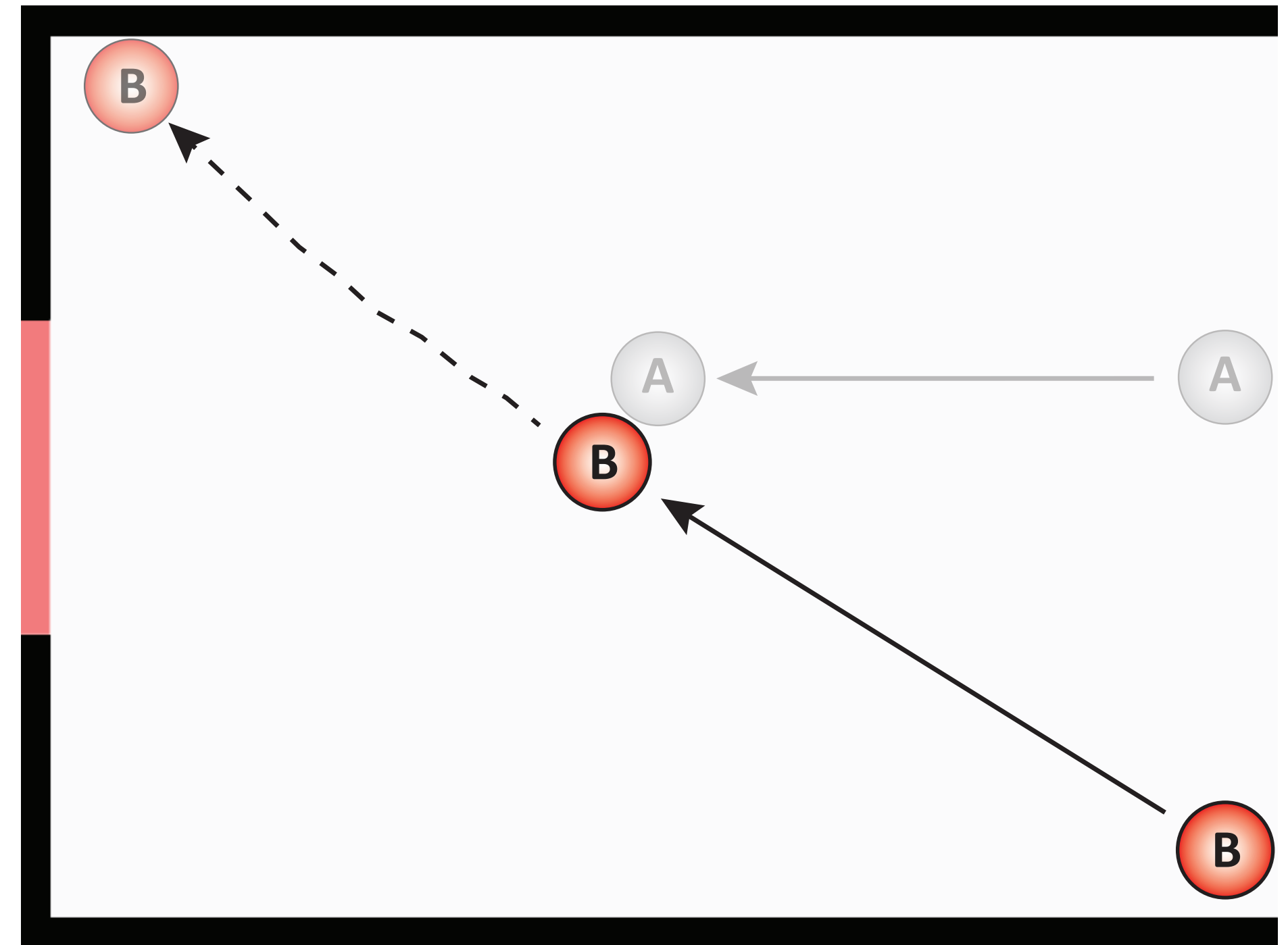
What happened?



Actual situation

 went through the gate

What would have happened?

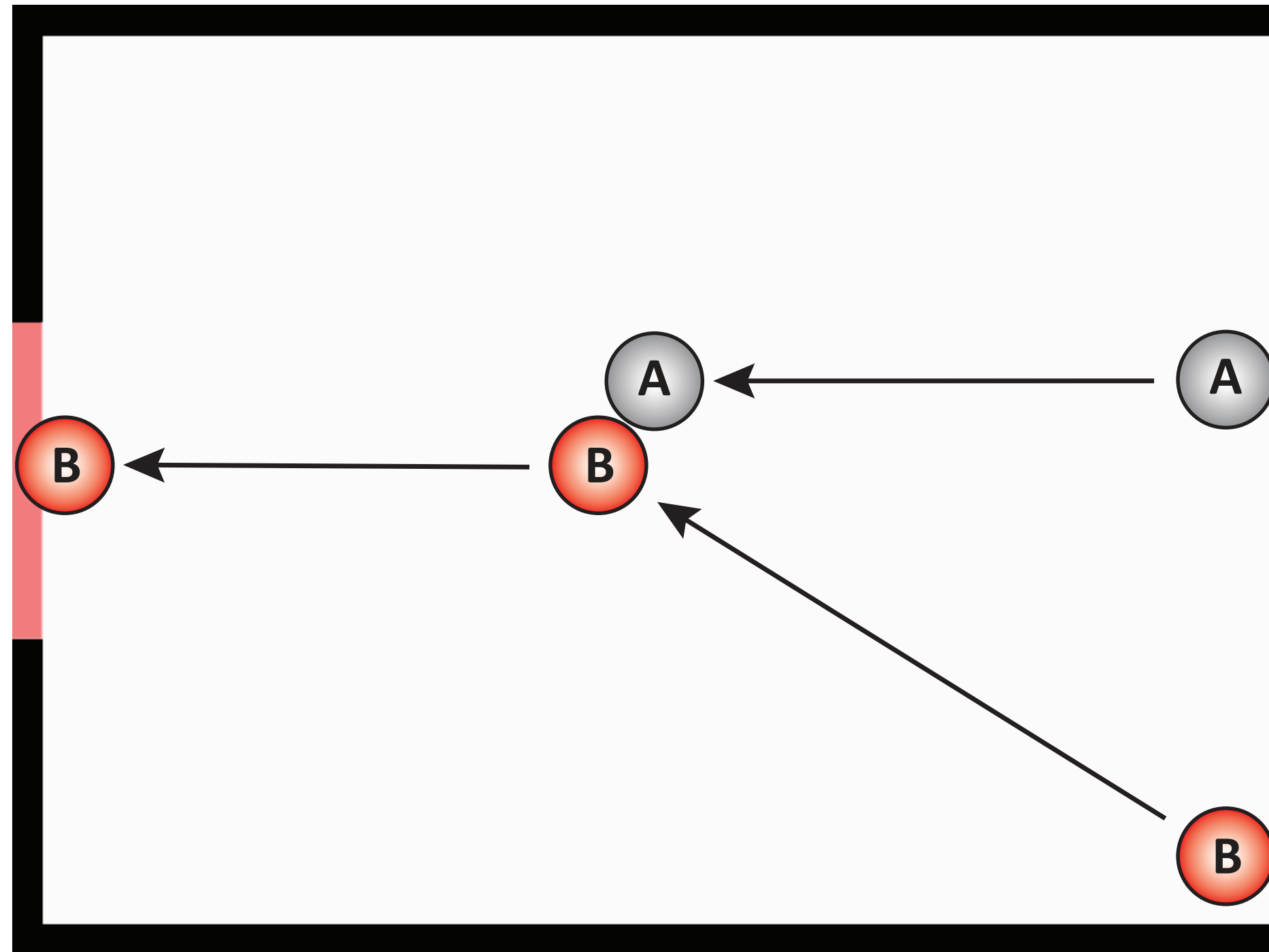


Counterfactual situation

 would have missed the gate 

\neq

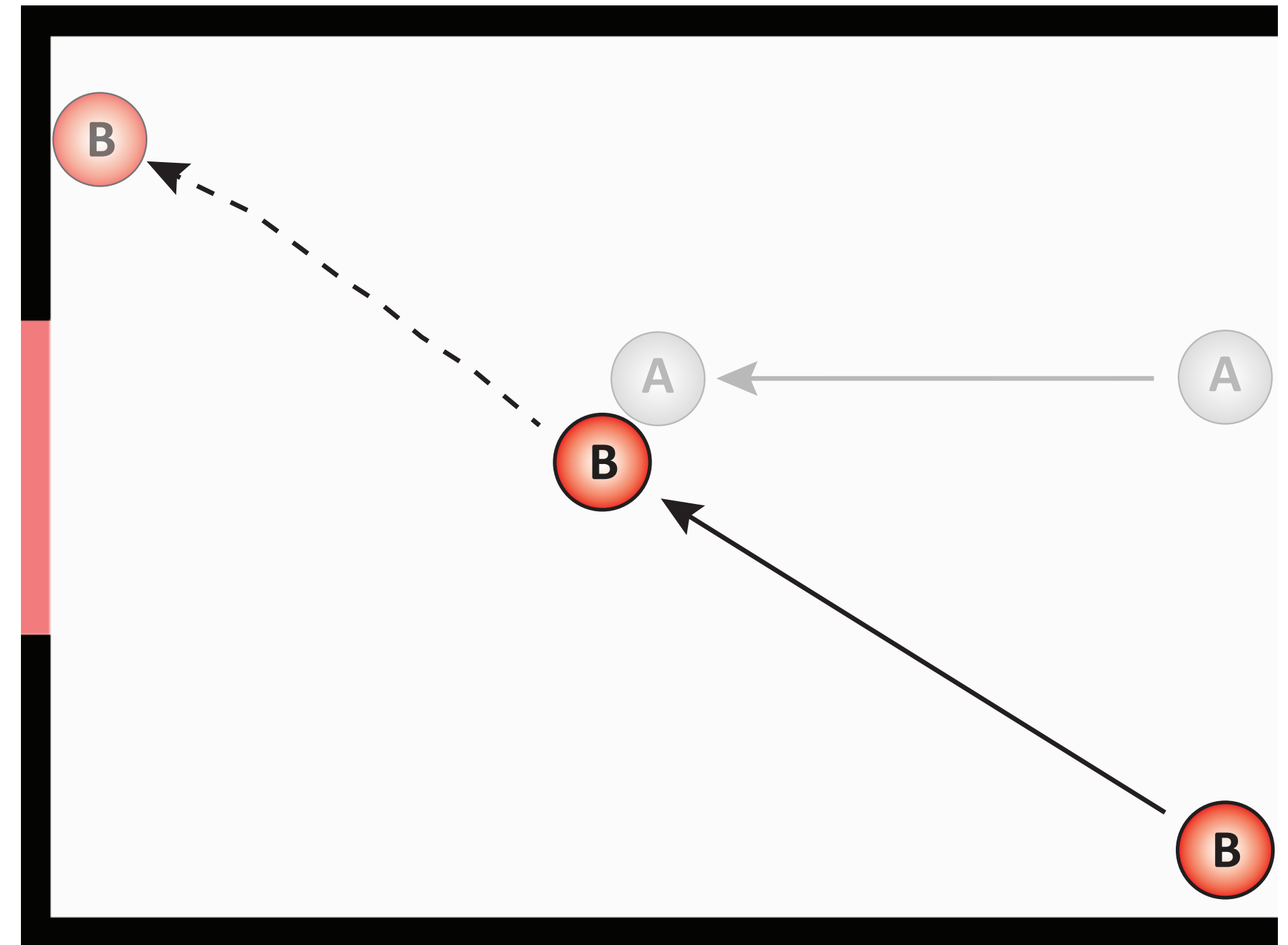
What happened?



Actual situation

B went through the gate

What would have happened?

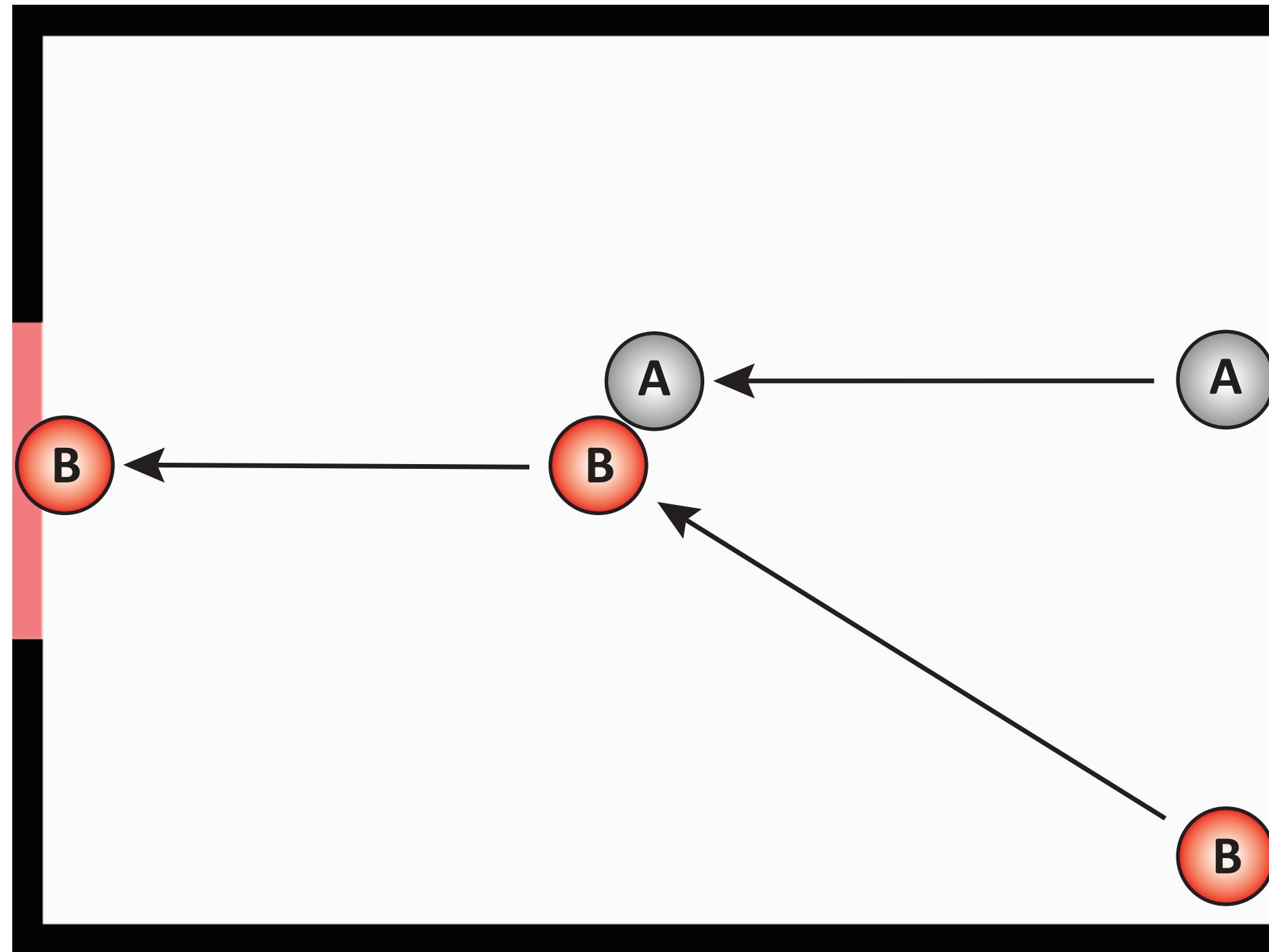


Counterfactual situation

B would have missed the gate ✓
B would have missed the gate ✓

\neq

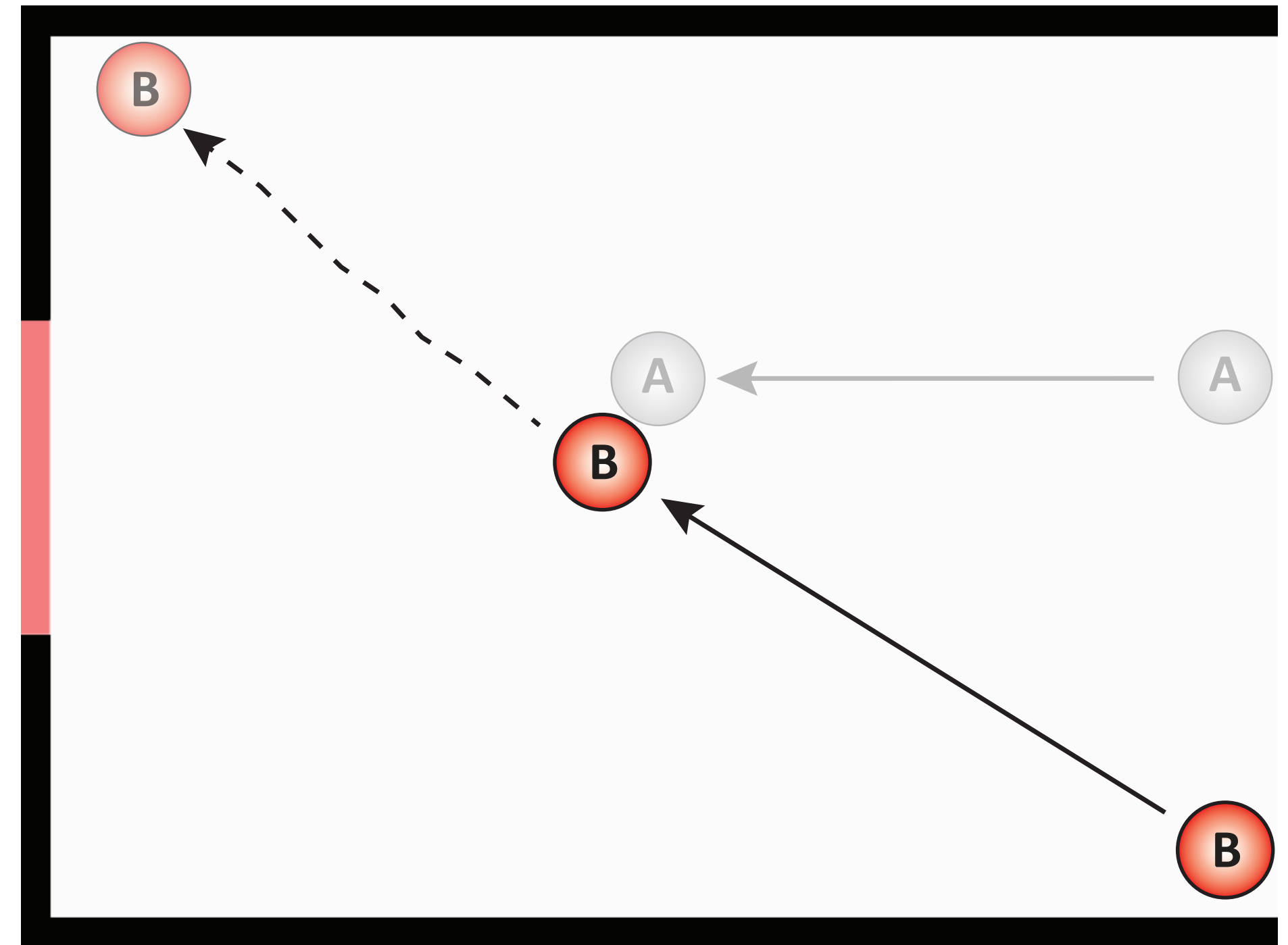
What happened?



Actual situation

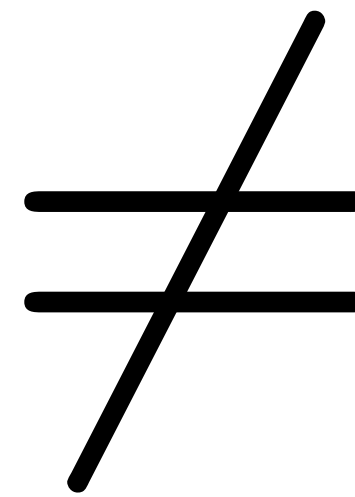
B went through the gate

What would have happened?

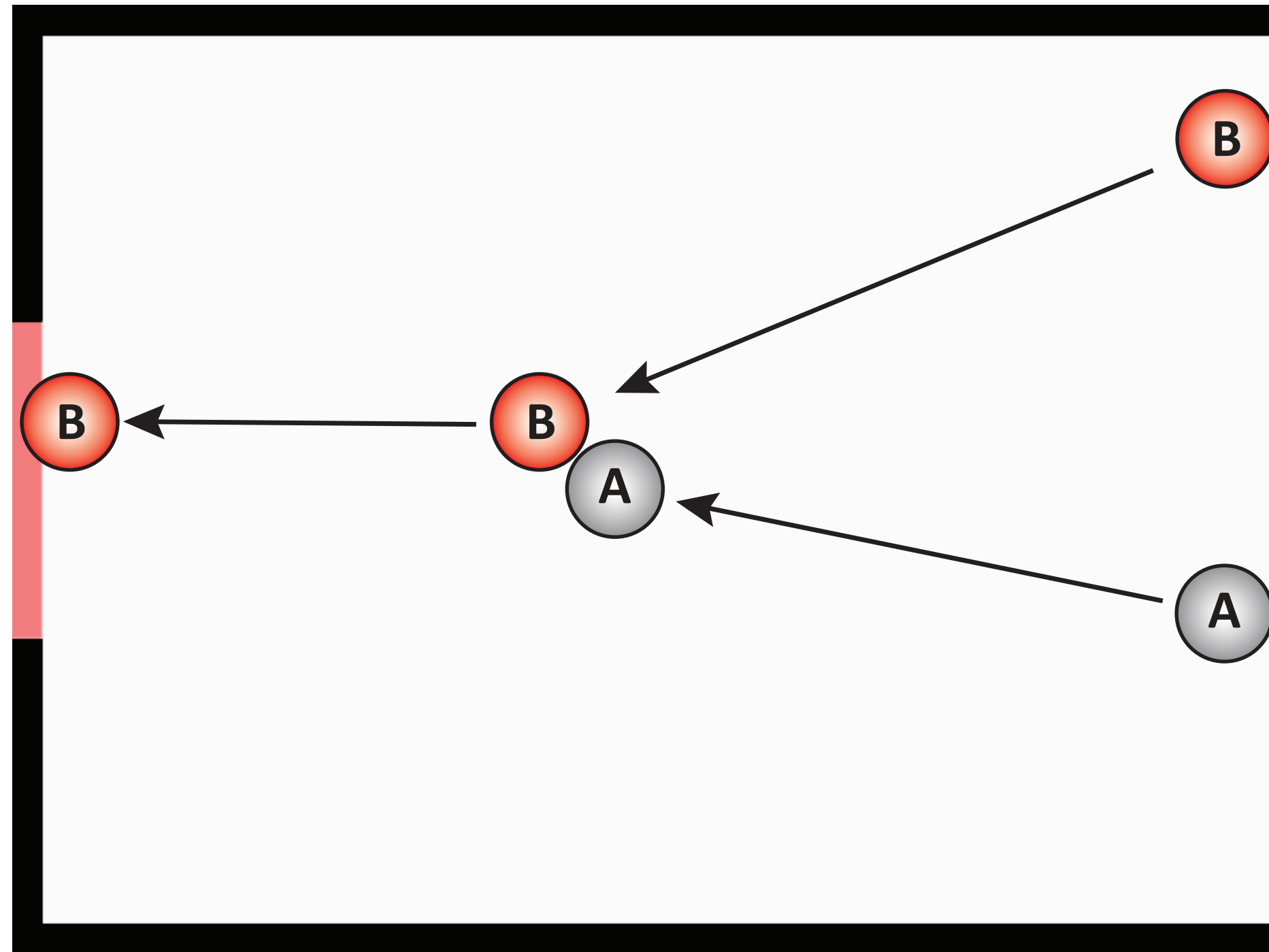


Counterfactual situation

- B** would have missed the gate ✓
- B** would have missed the gate ✓
- B** would have missed the gate ✓



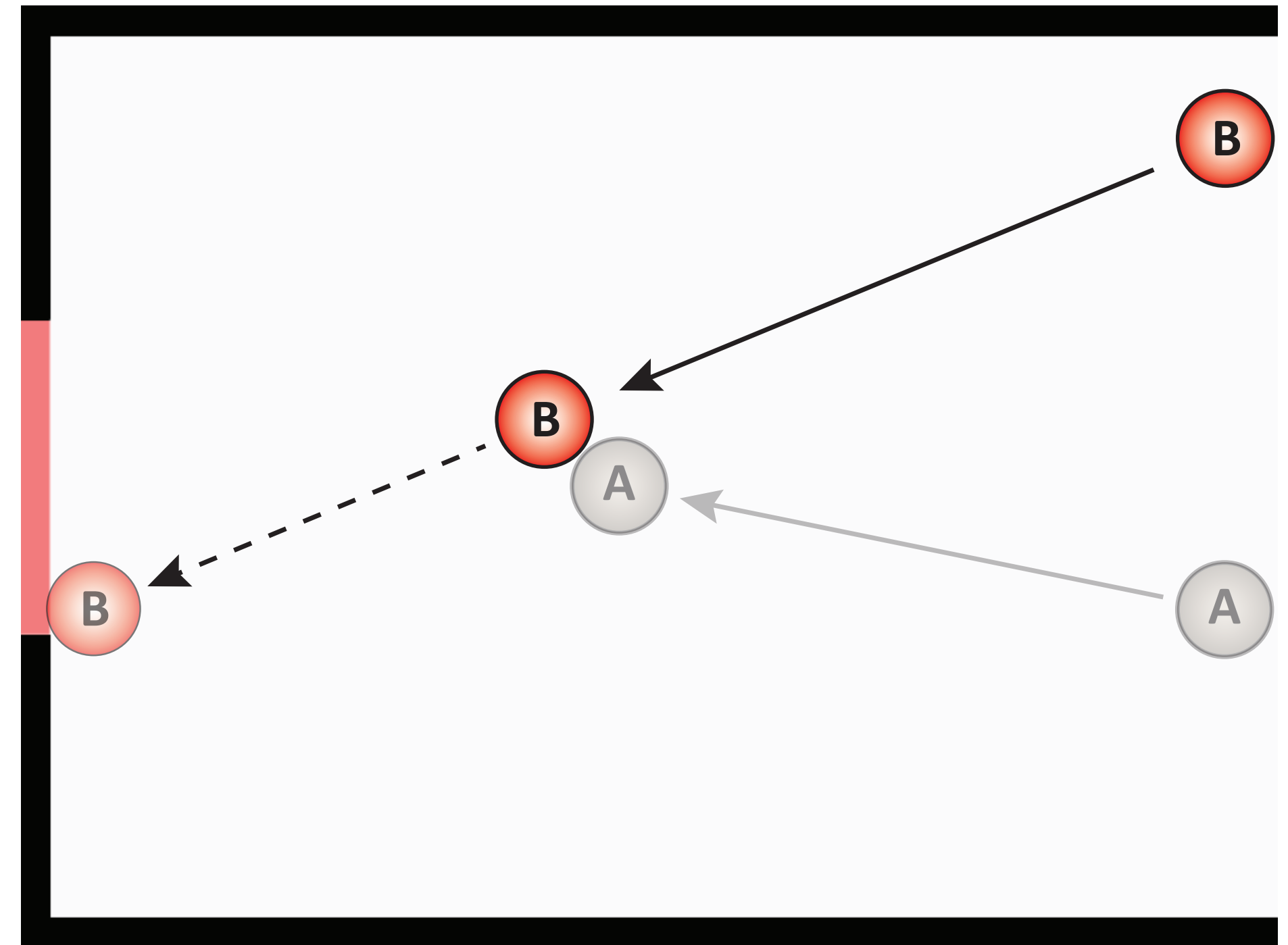
What happened?



Actual situation

 went through the gate

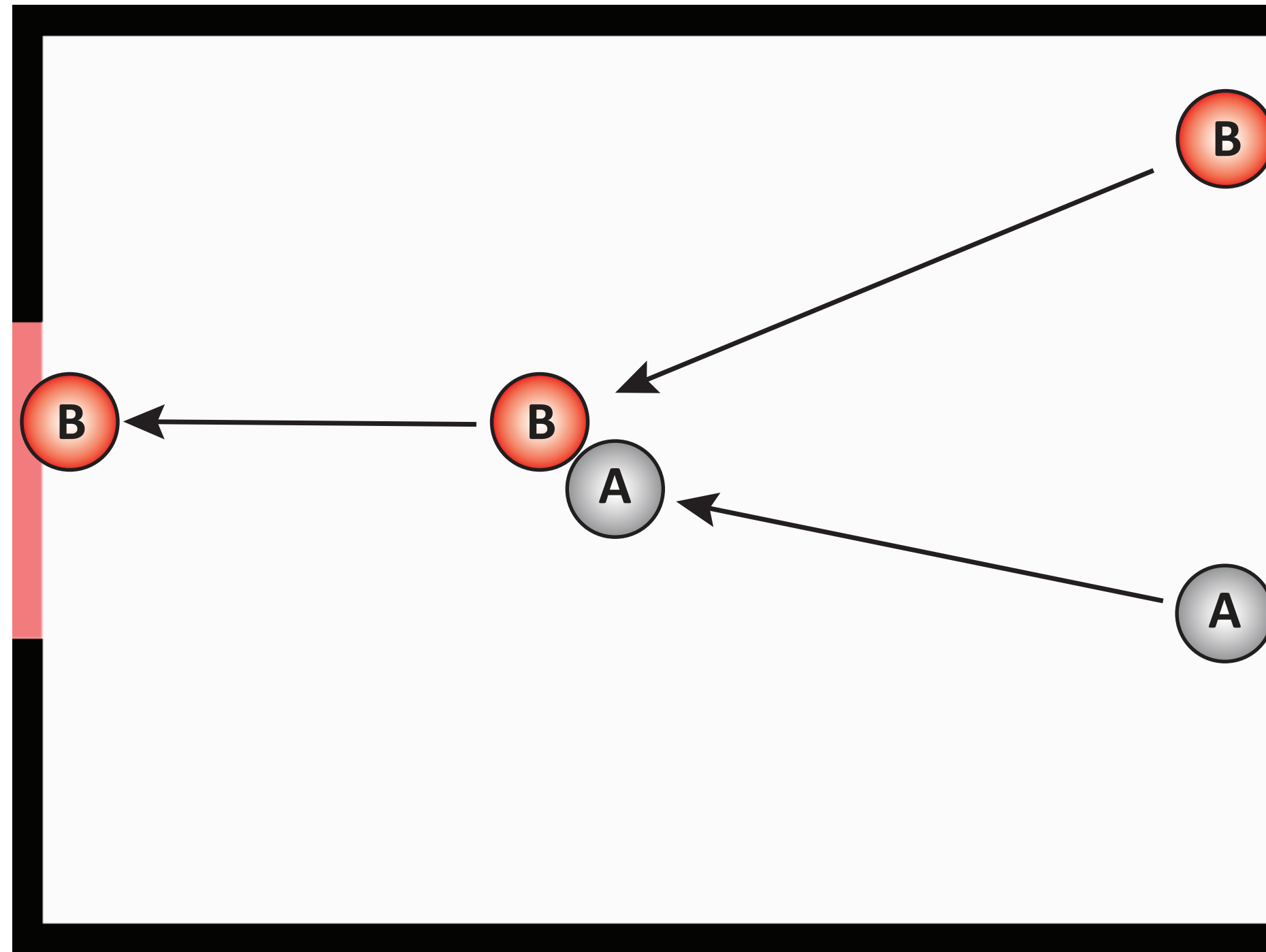
What would have happened?



Counterfactual situation

\neq

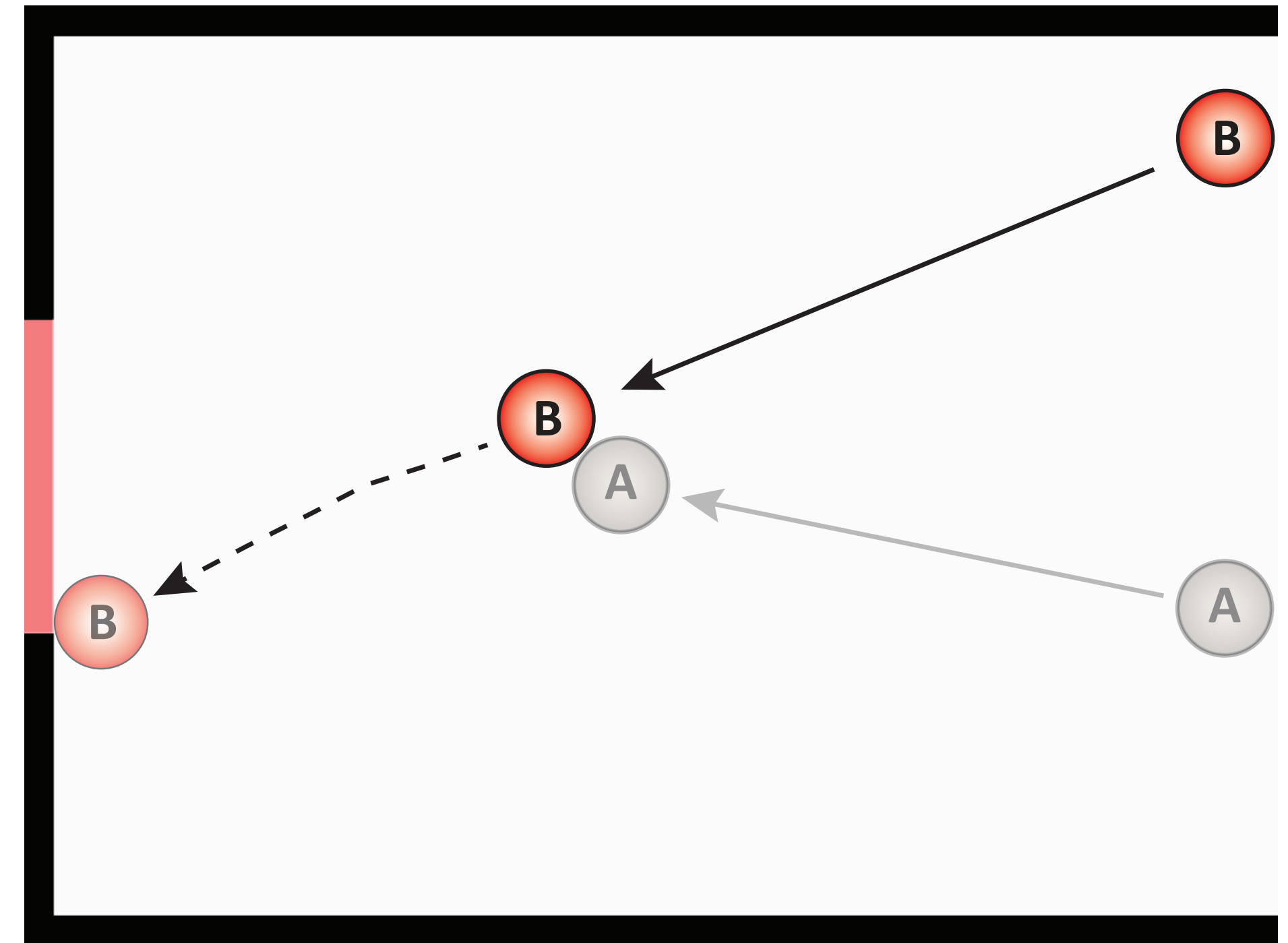
What happened?



Actual situation

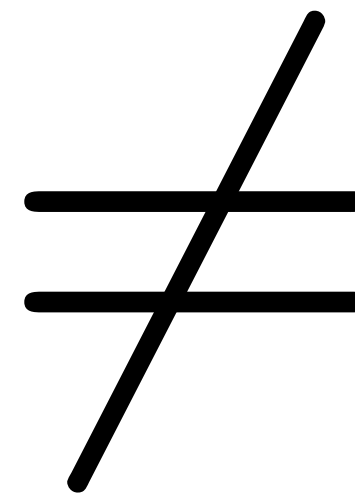
B went through the gate

What would have happened?

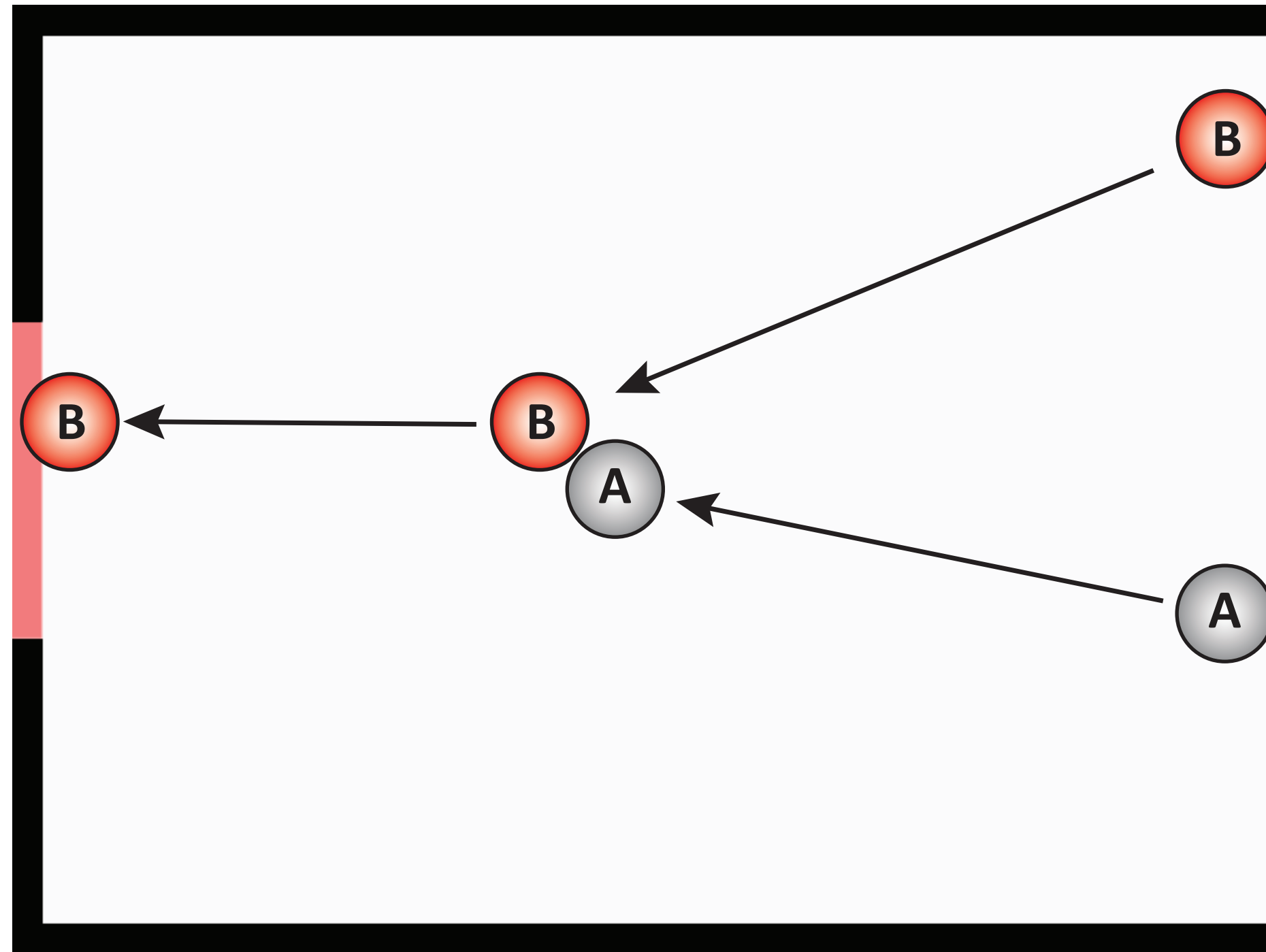


Counterfactual situation

B would have missed the gate ✓



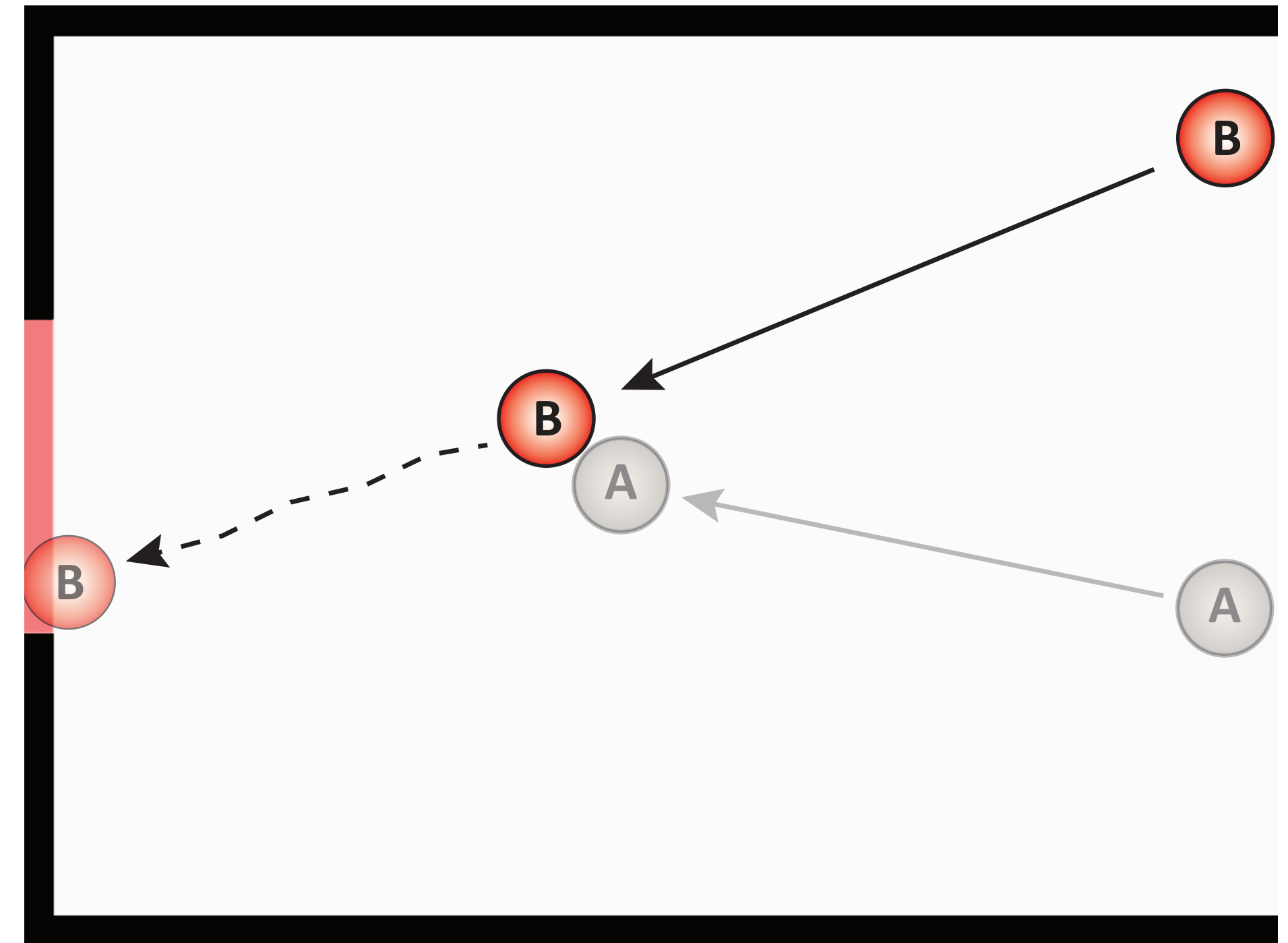
What happened?



Actual situation

B went through the gate

What would have happened?

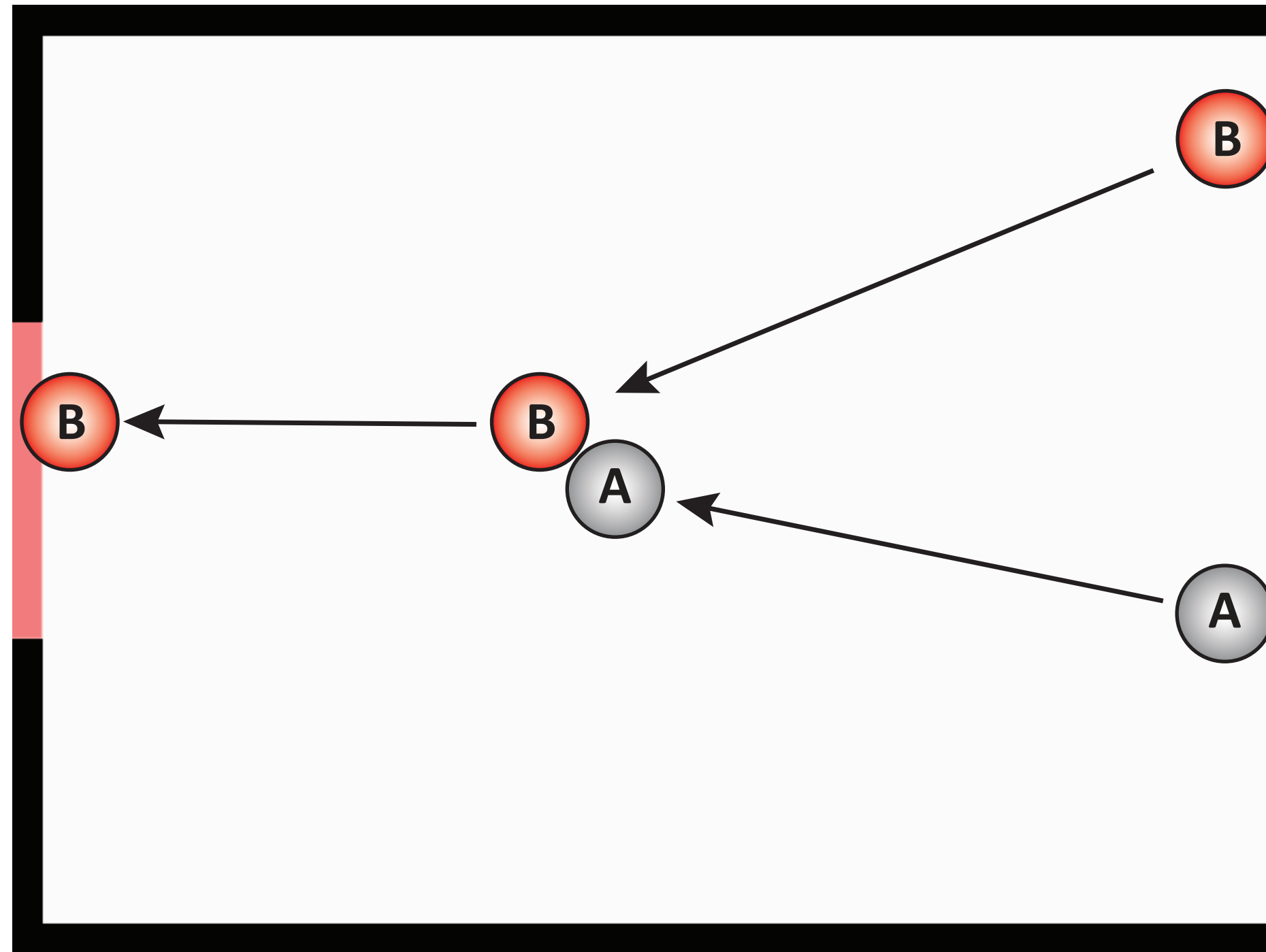


Counterfactual situation

B would have missed the gate ✓
B would have missed the gate ✗

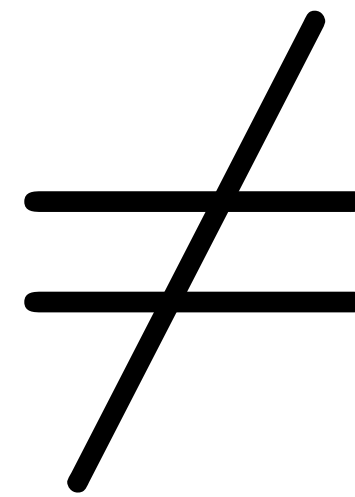
≠

What happened?

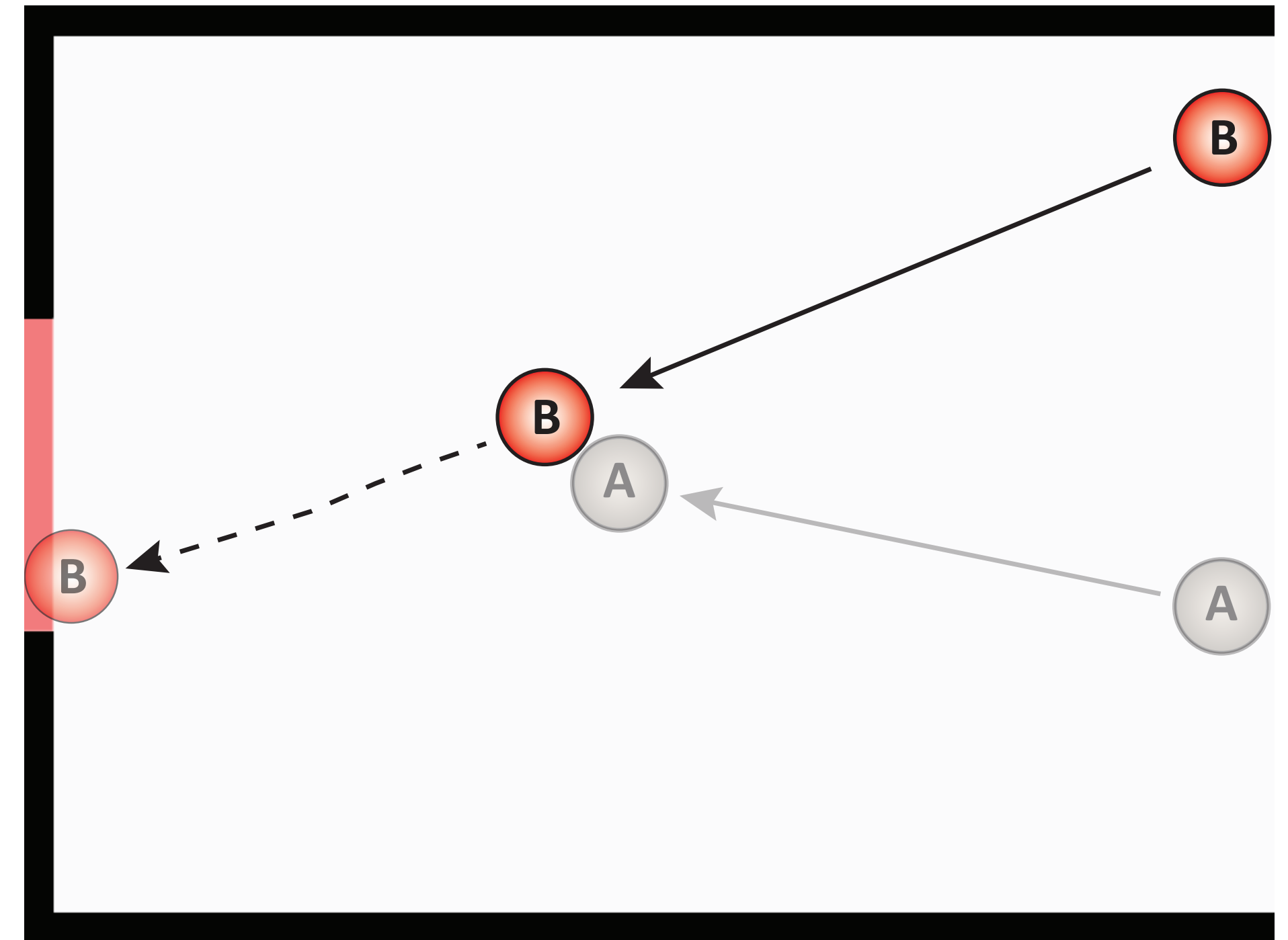


Actual situation

B went through the gate

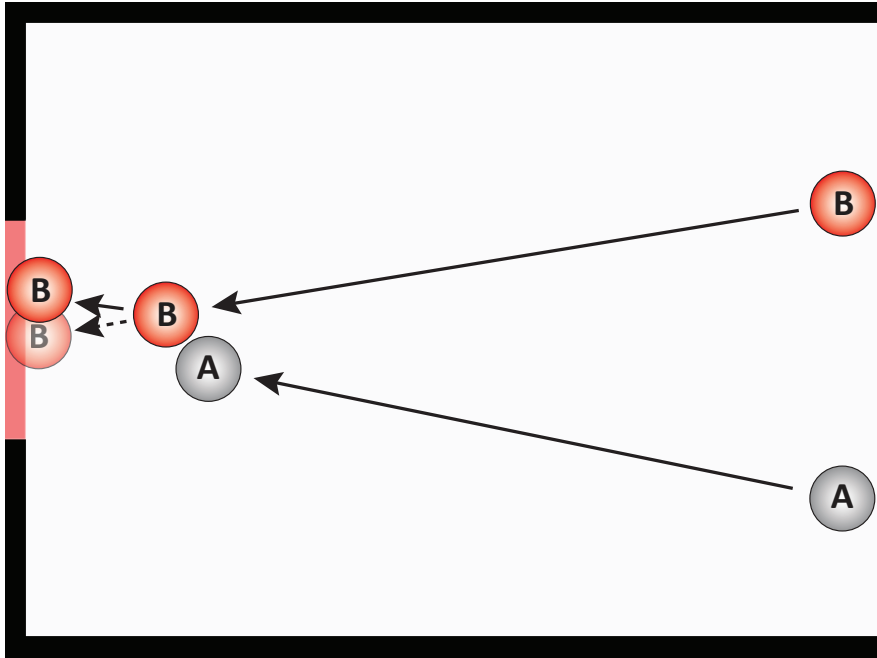
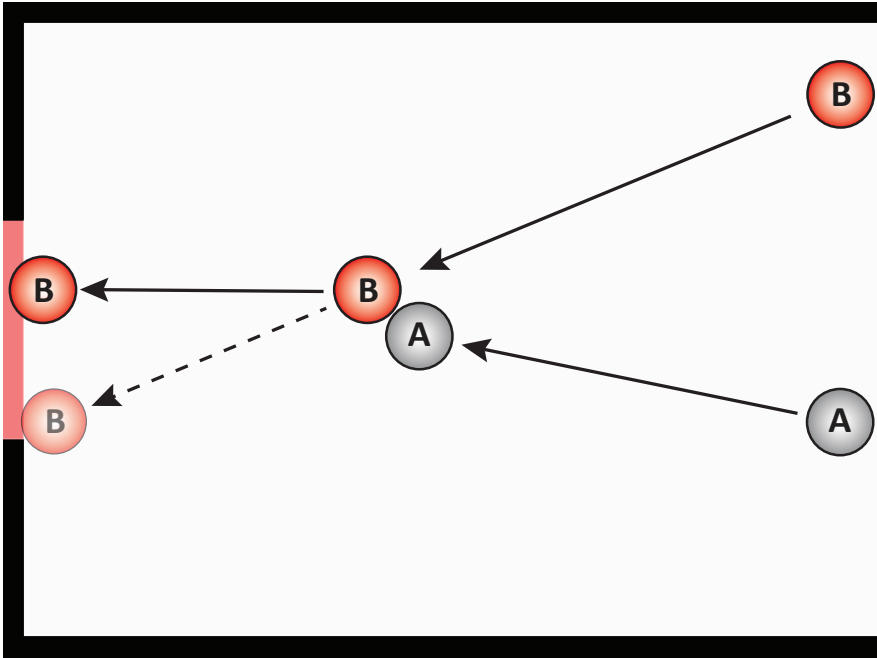
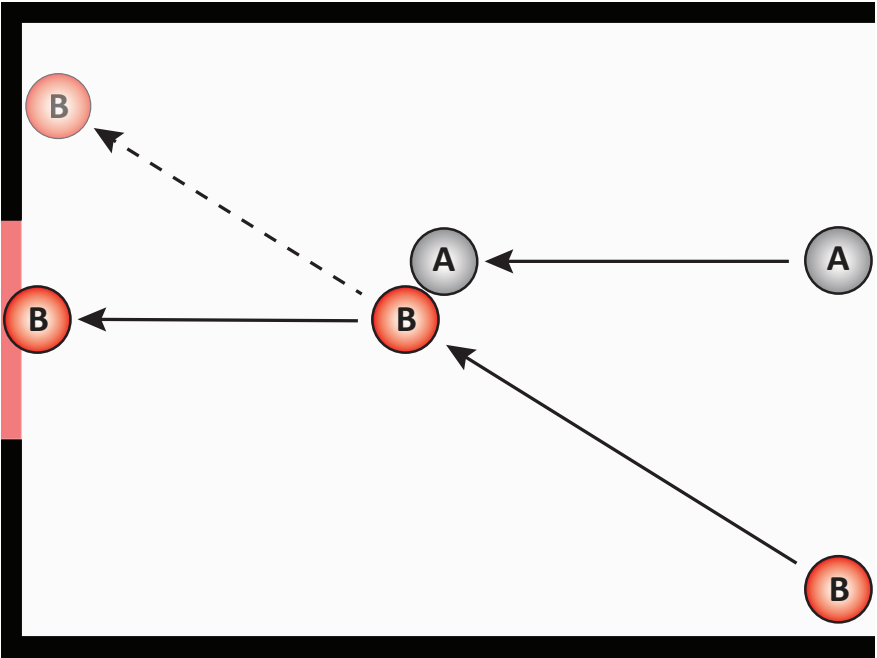


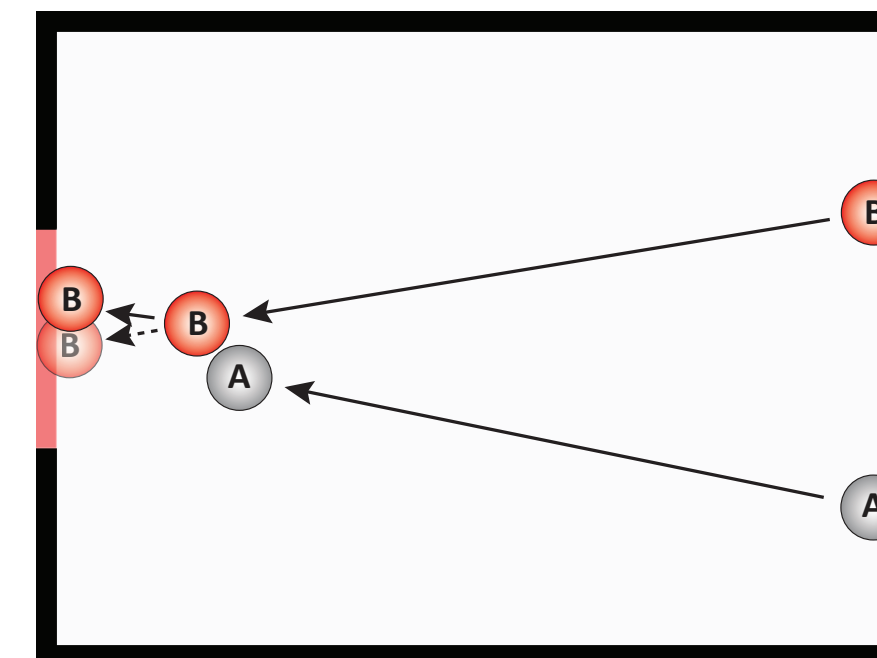
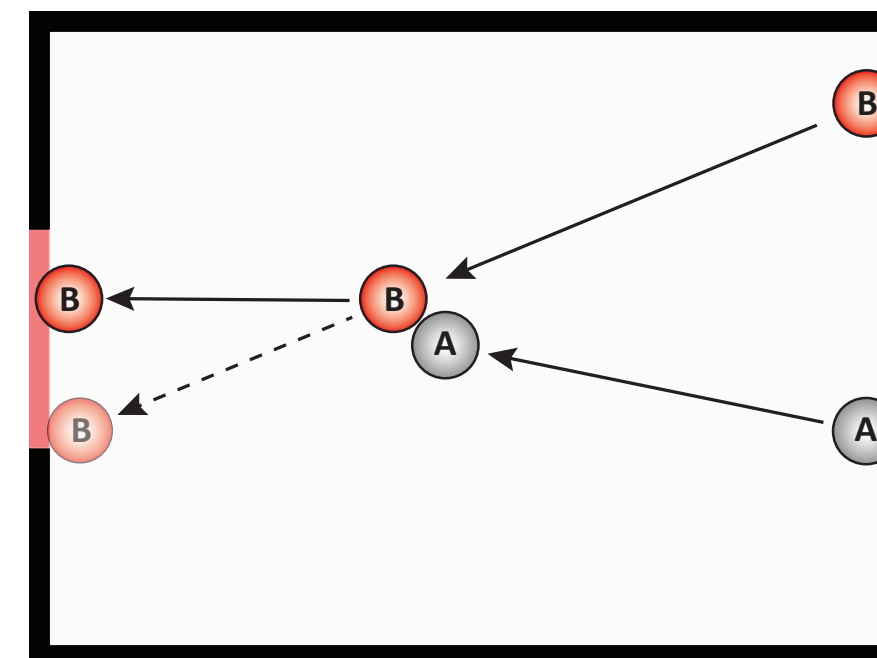
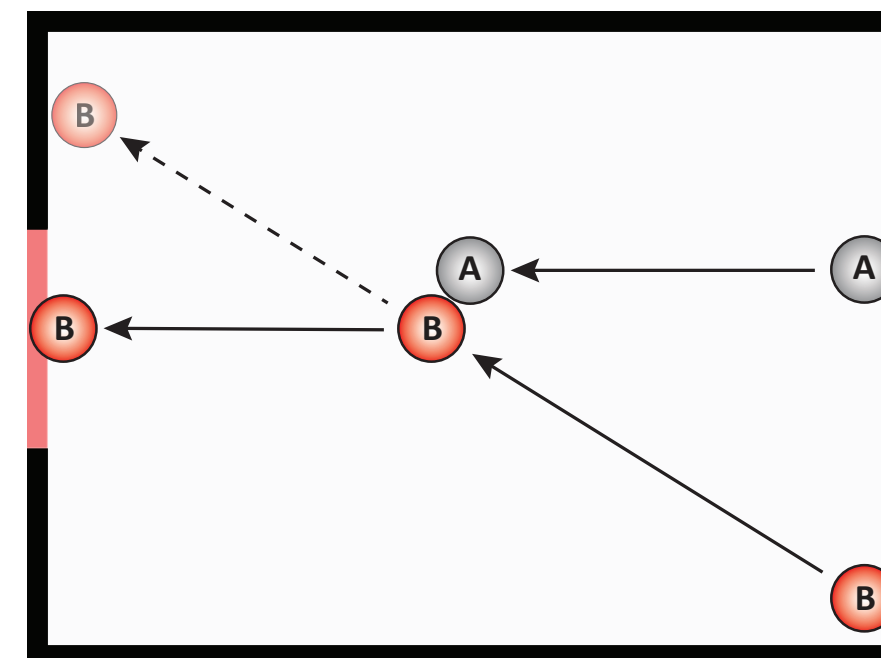
What would have happened?



Counterfactual situation

- B** would have missed the gate ✓
- B** would have missed the gate ✗
- B** would have missed the gate ✗

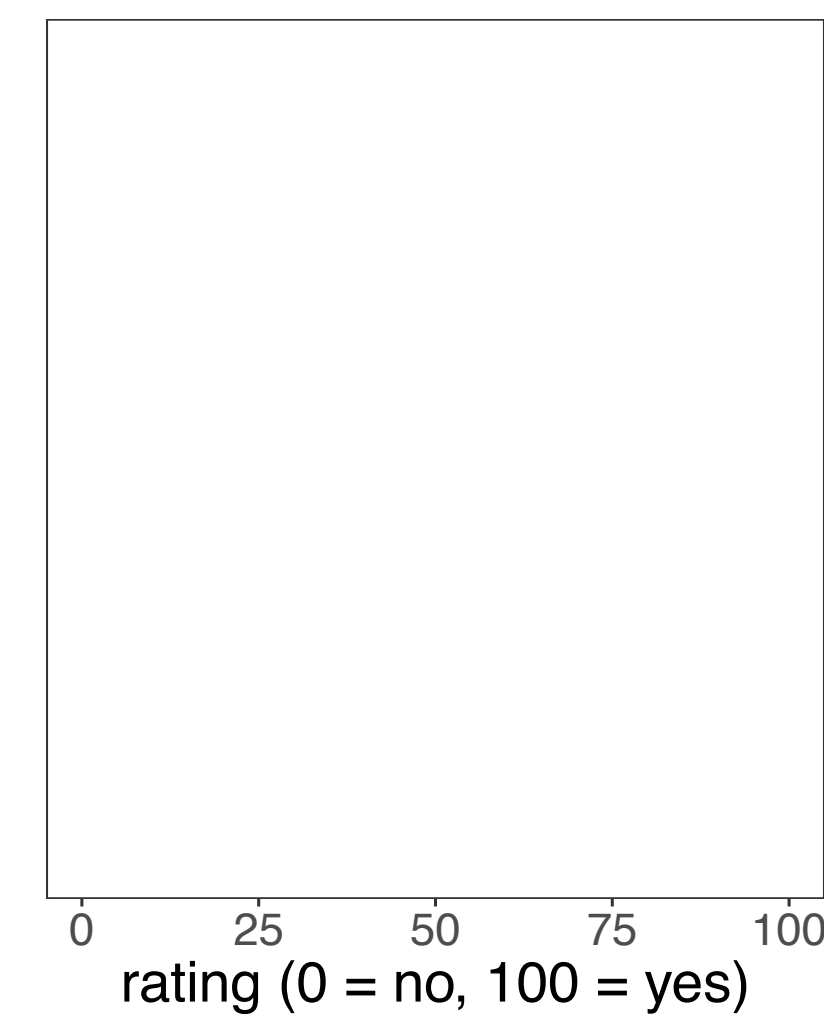
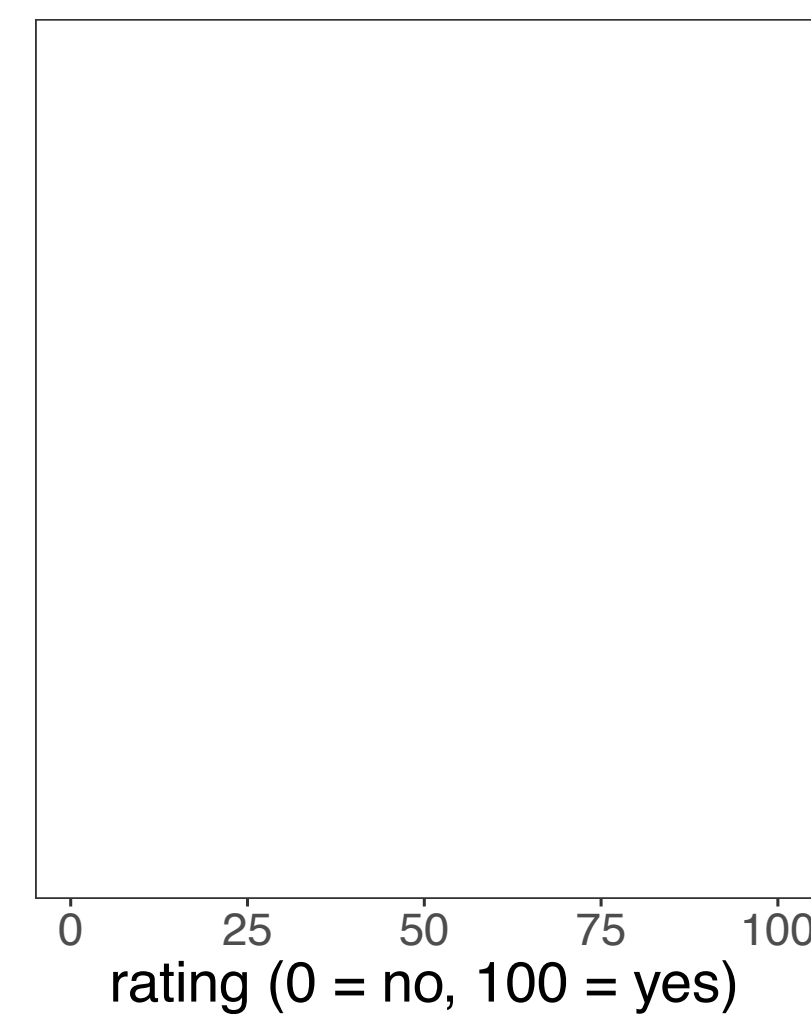
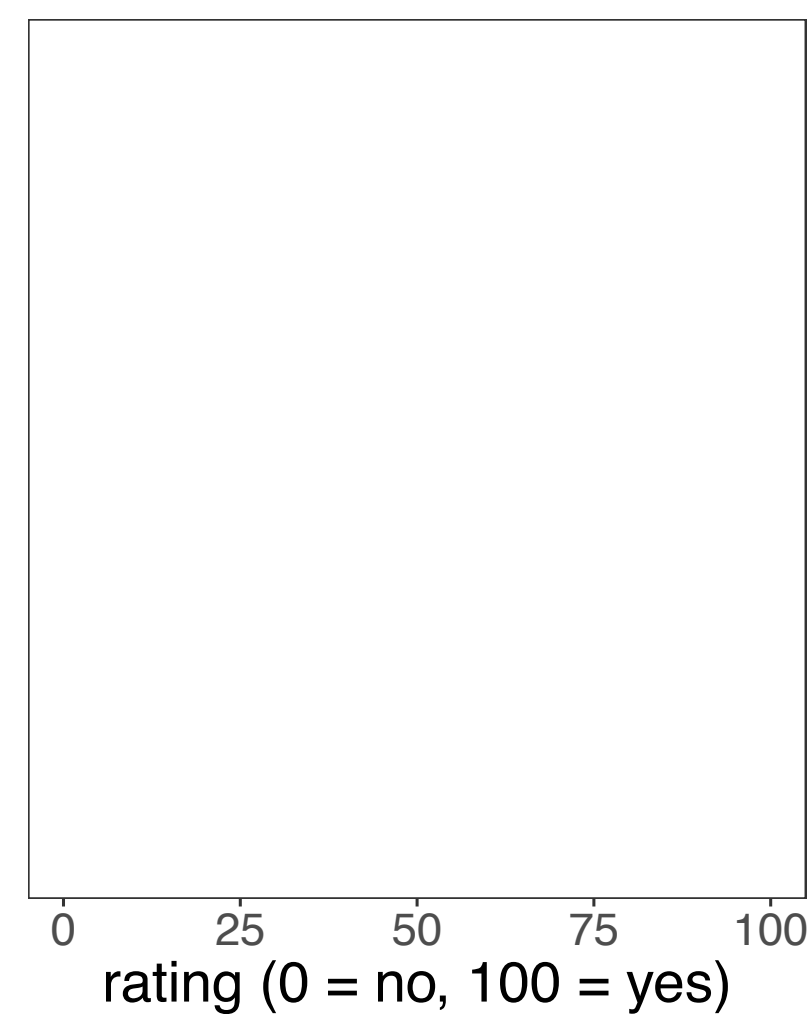


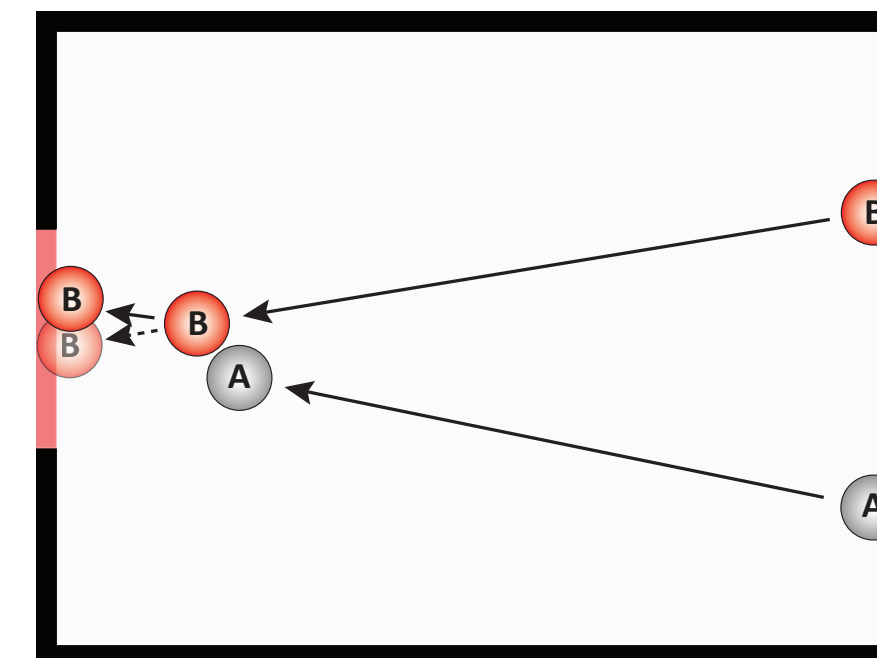
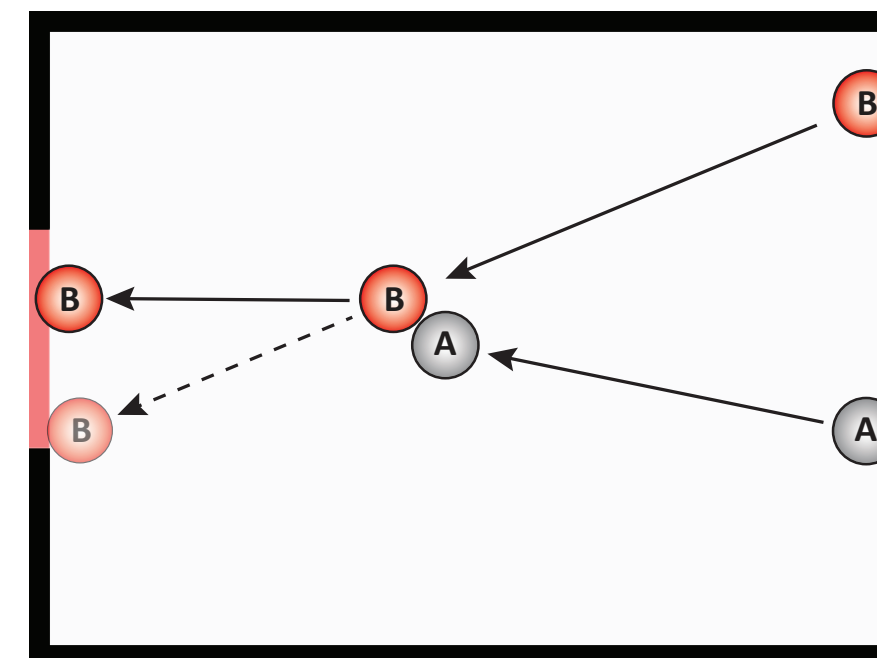
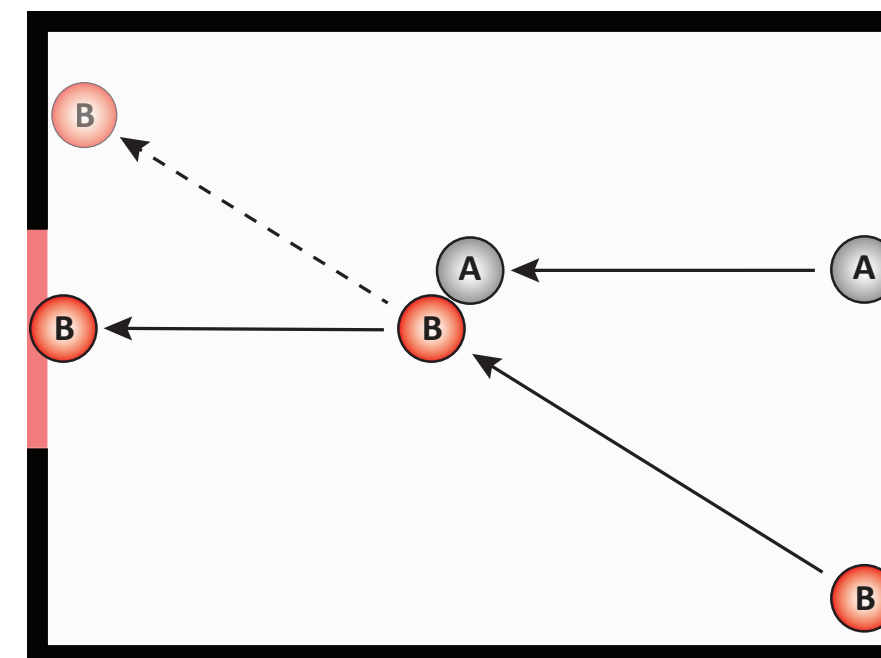


Counterfactual
simulation model

Did A cause B to go
through the gate?

Would B have
missed the gate?

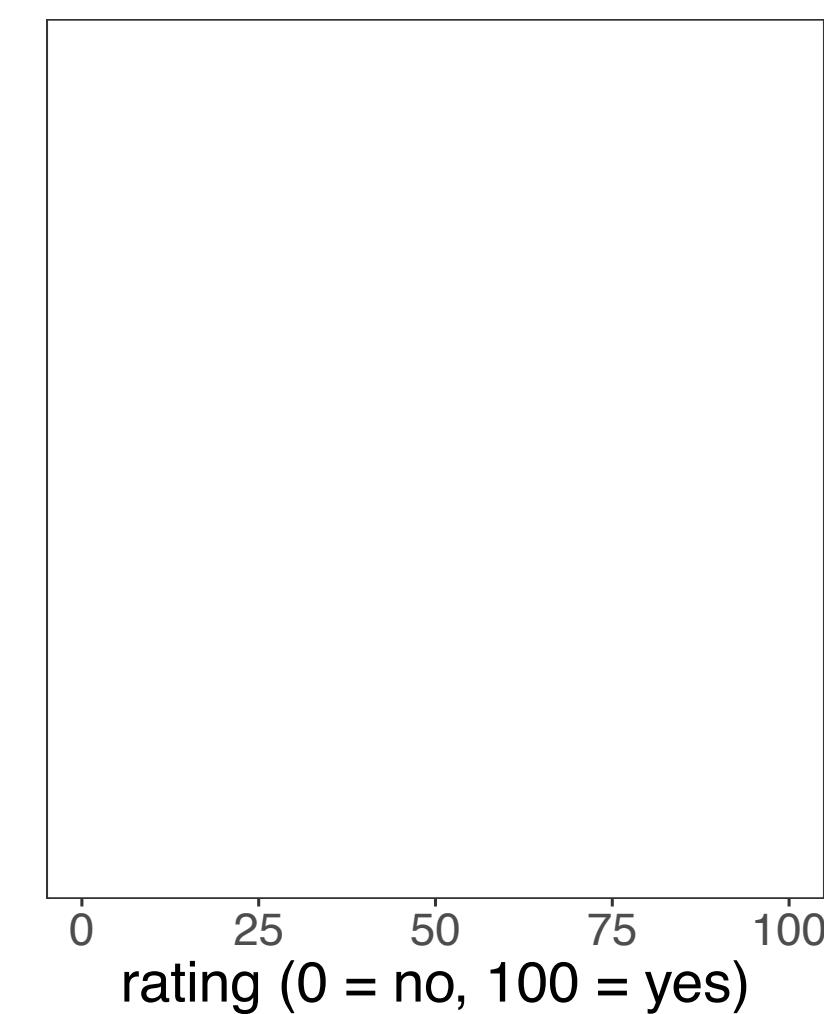
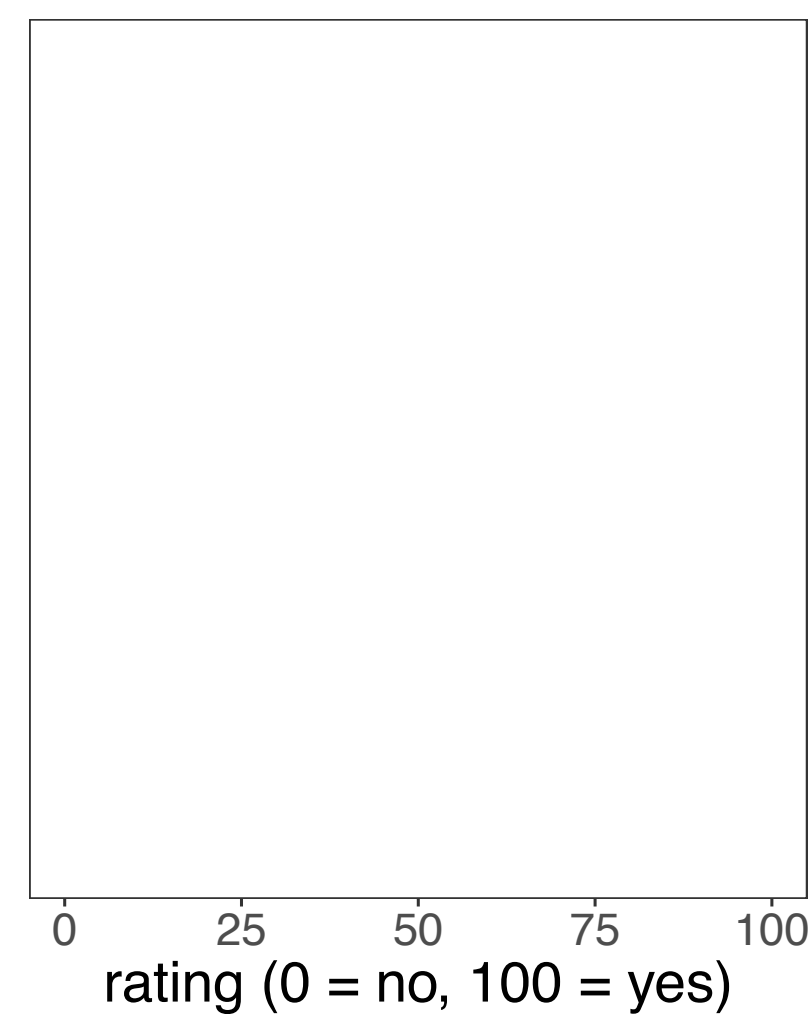
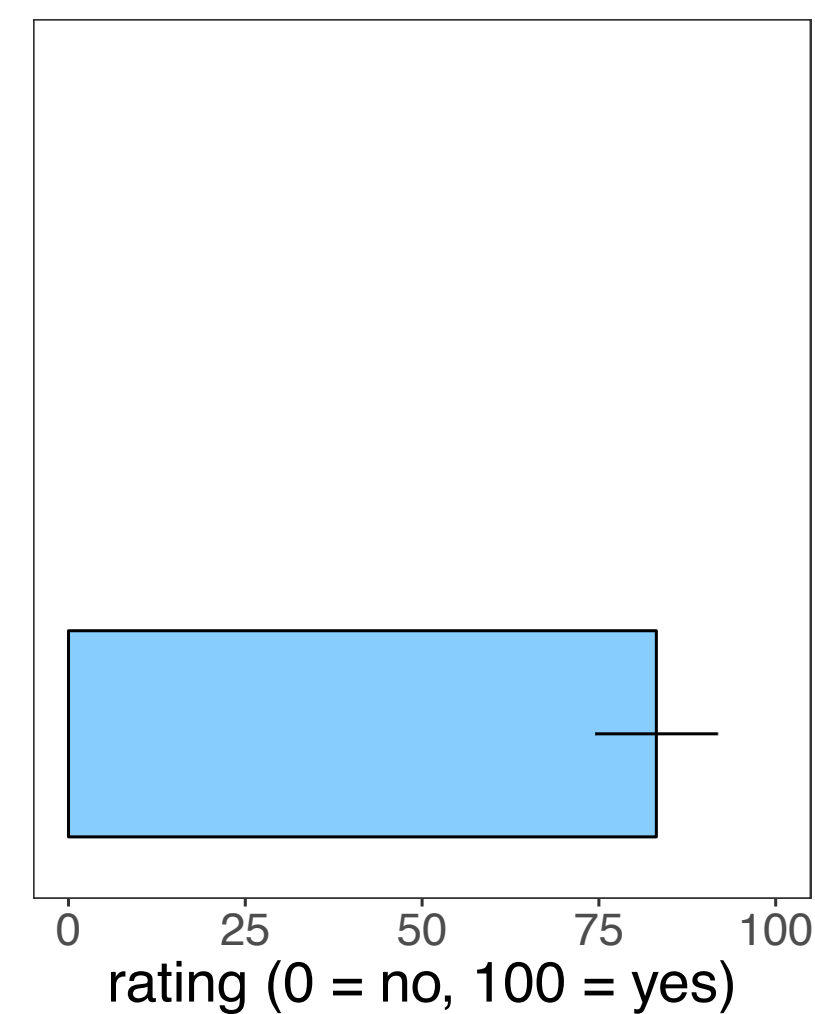


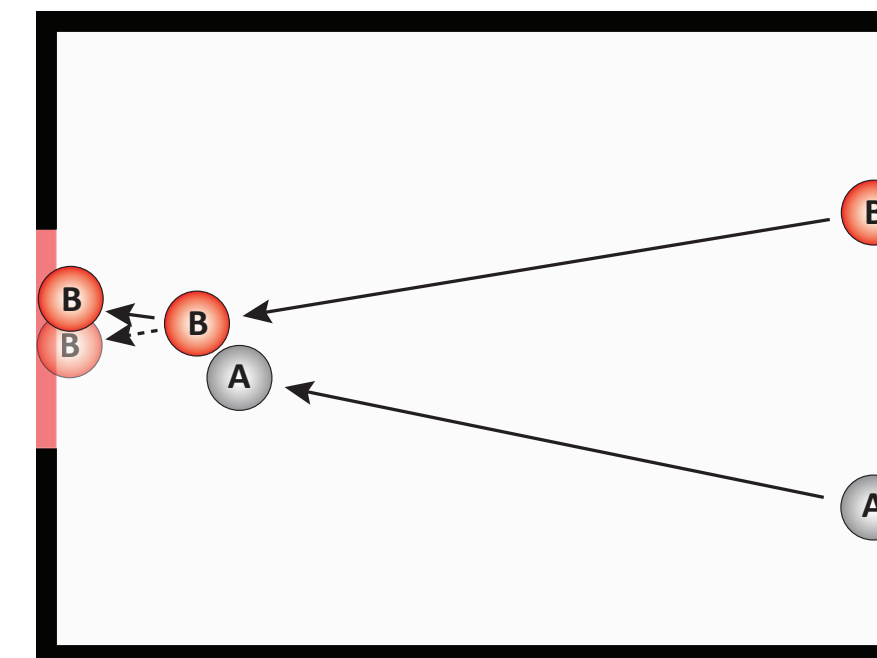
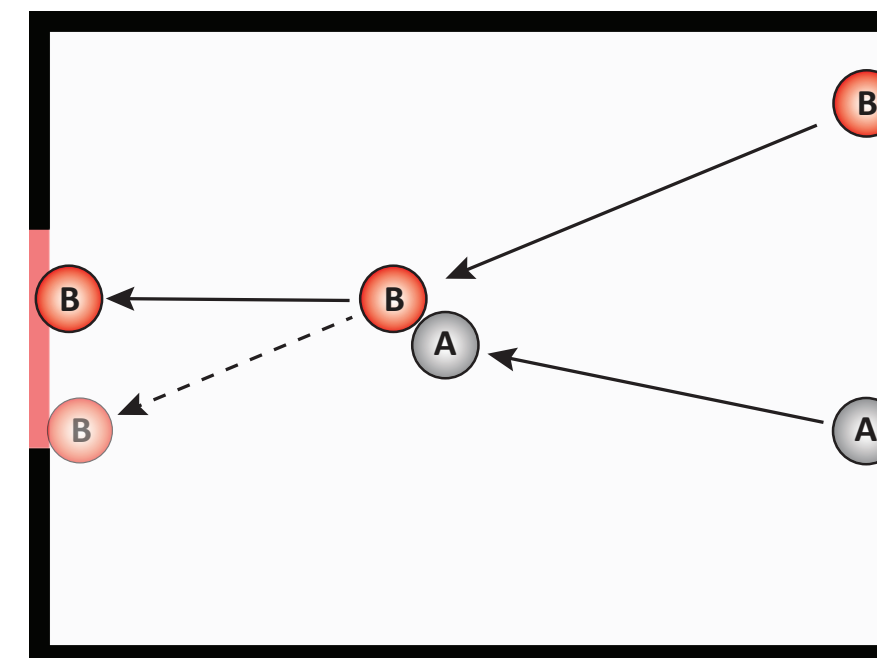
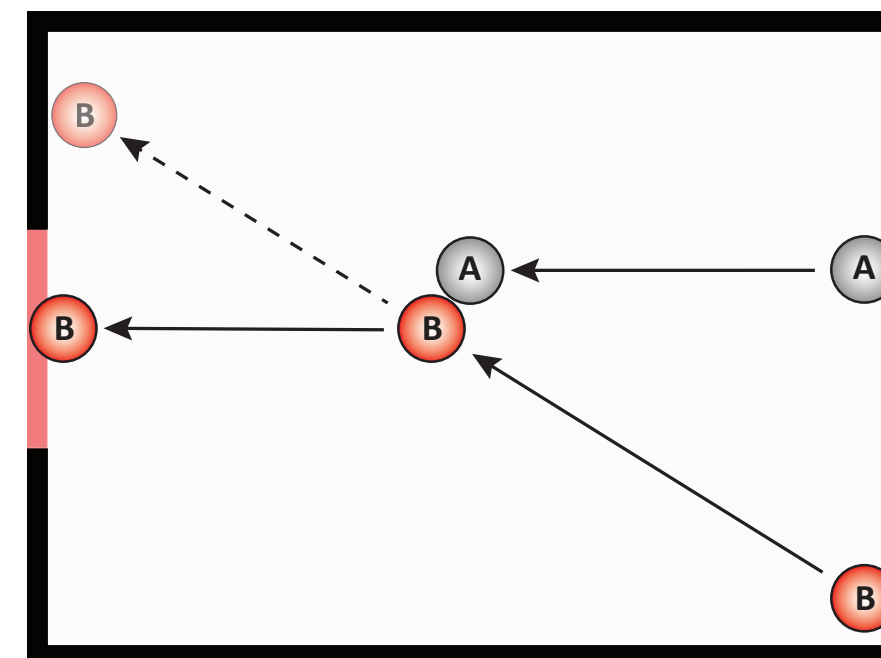


Counterfactual
simulation model

Did A cause B to go
through the gate?

Would B have
missed the gate?

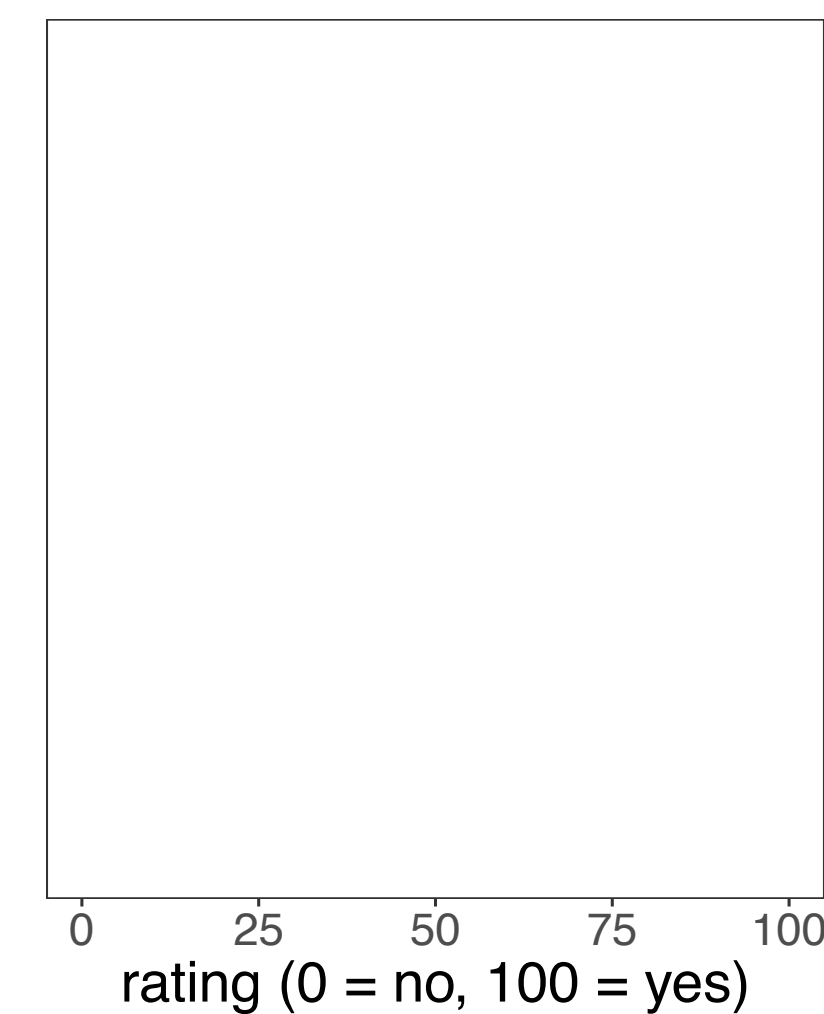
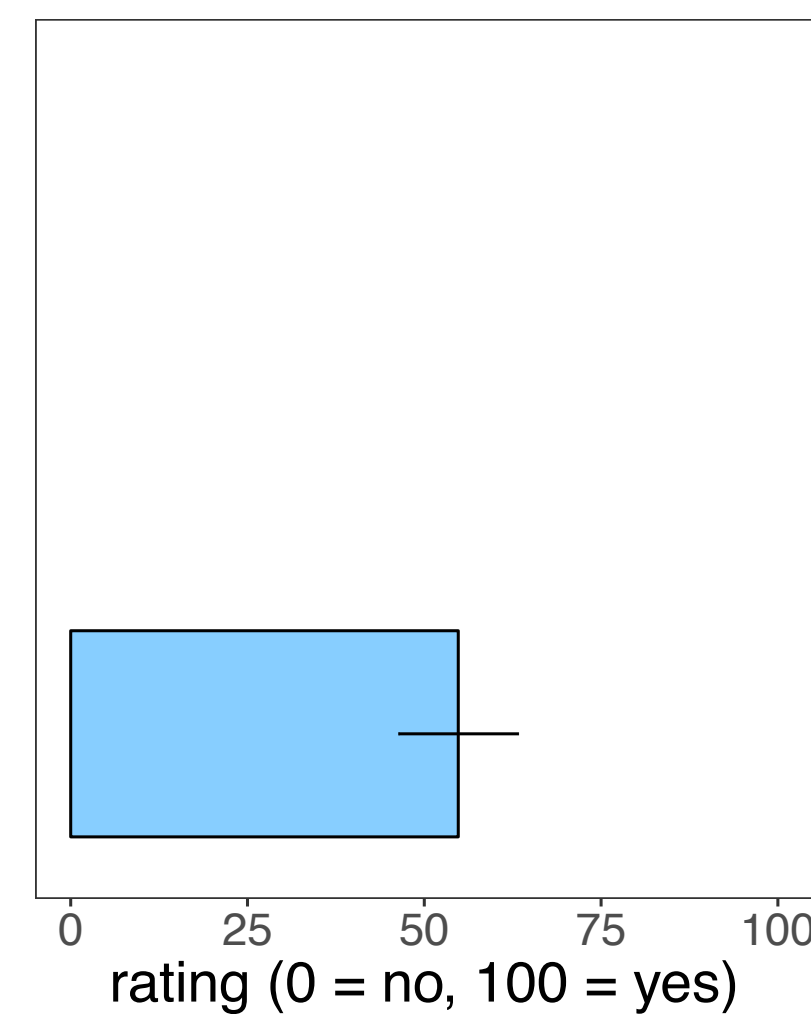
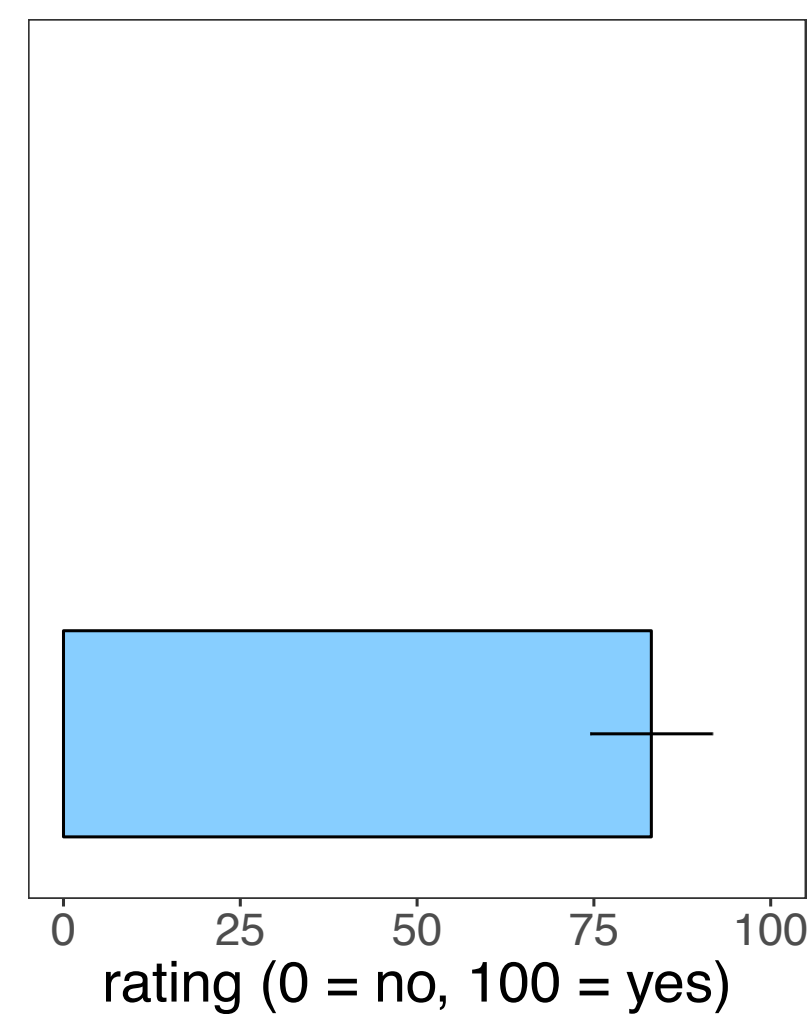


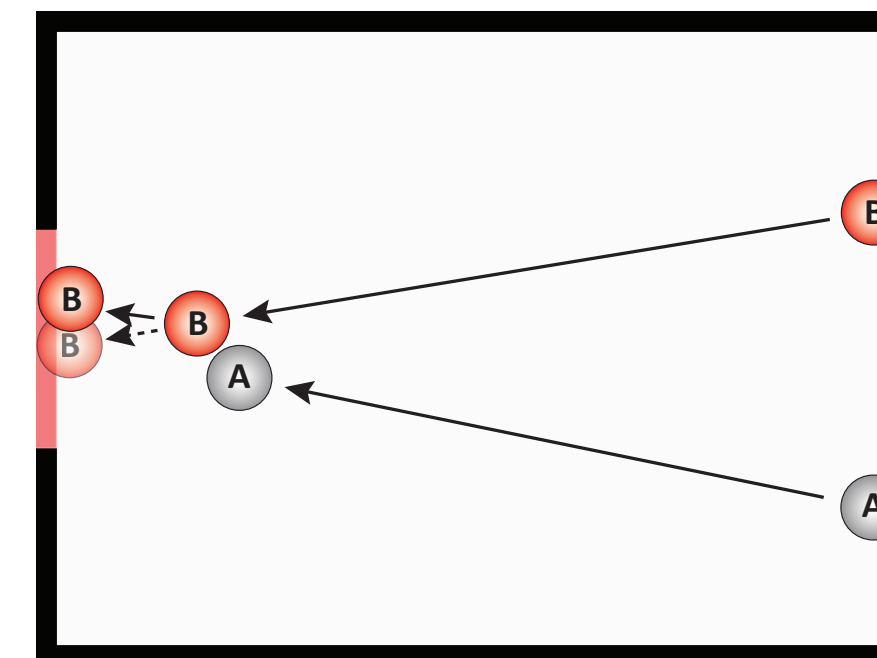
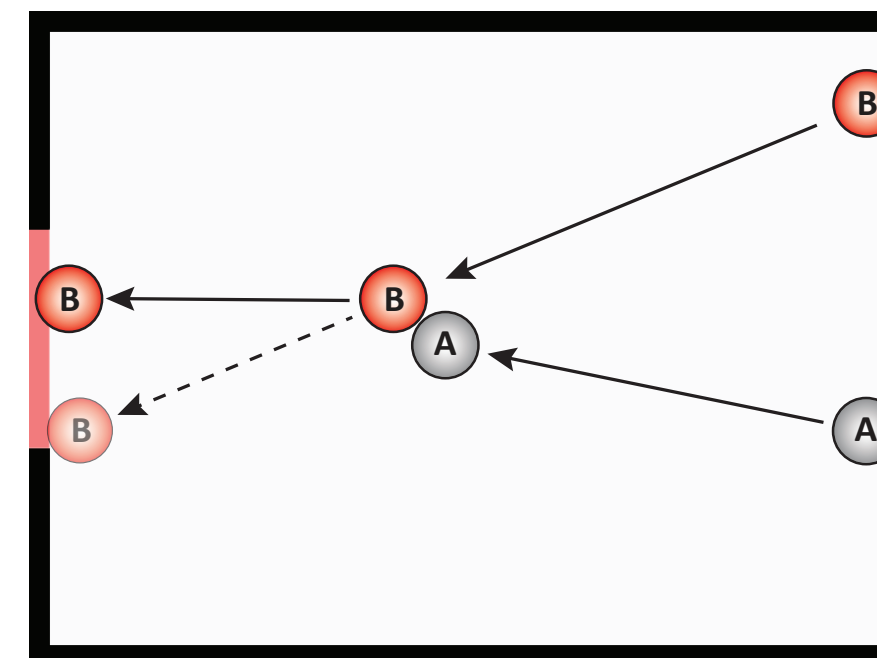
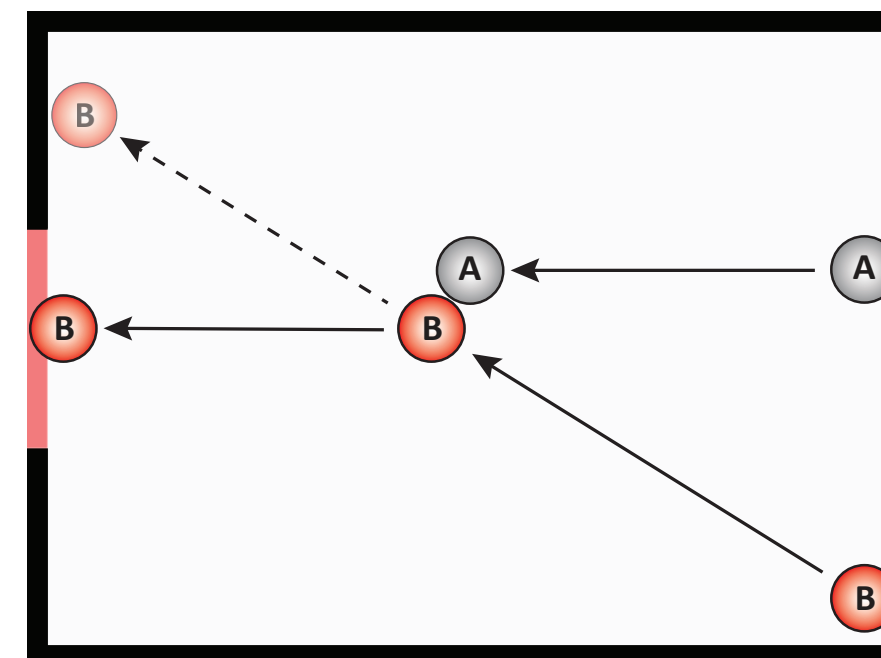


Counterfactual
simulation model

Did A cause B to go
through the gate?

Would B have
missed the gate?

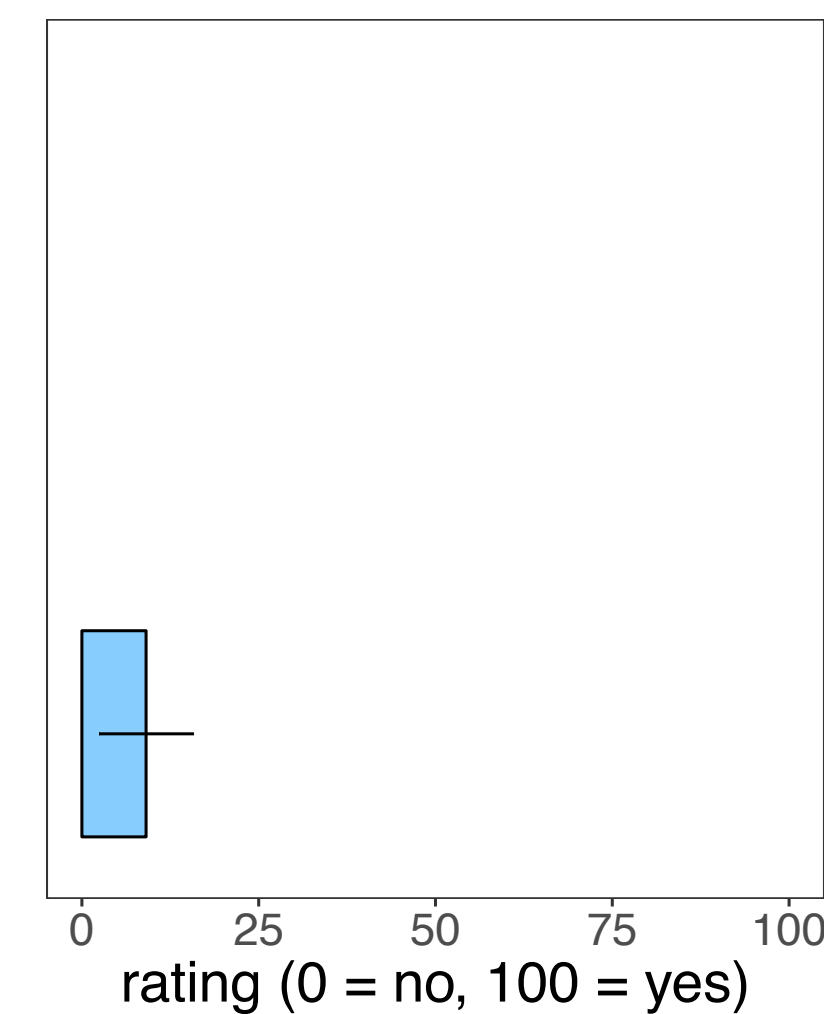
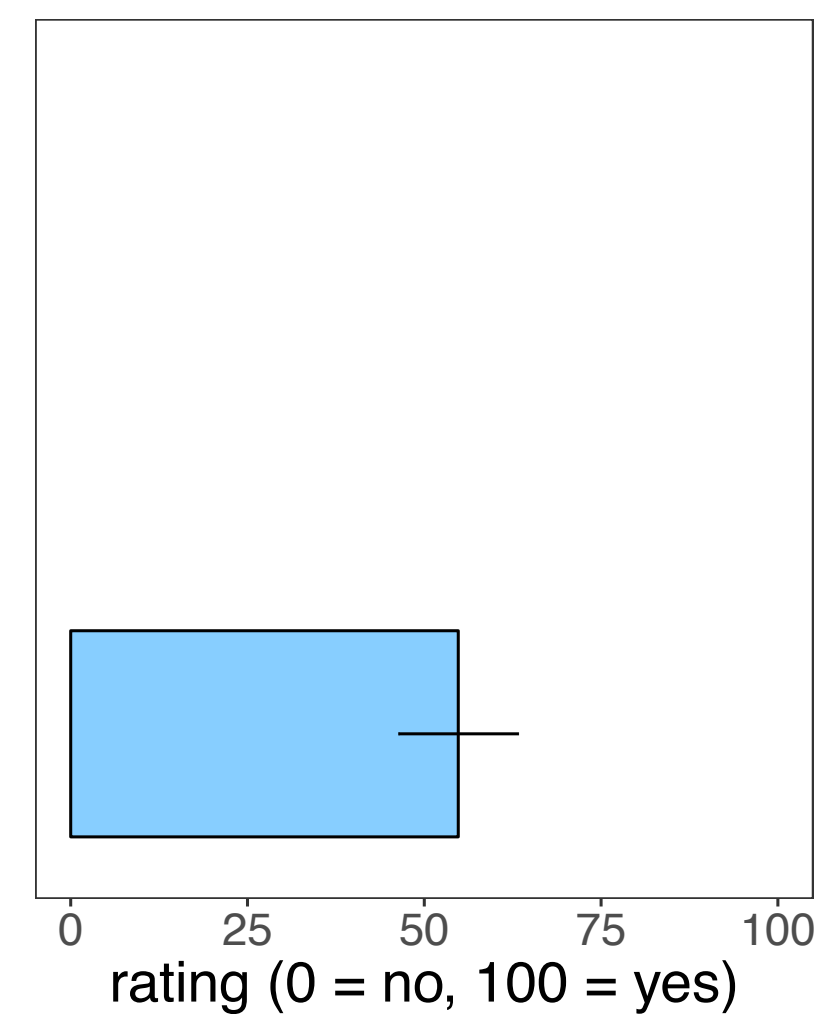
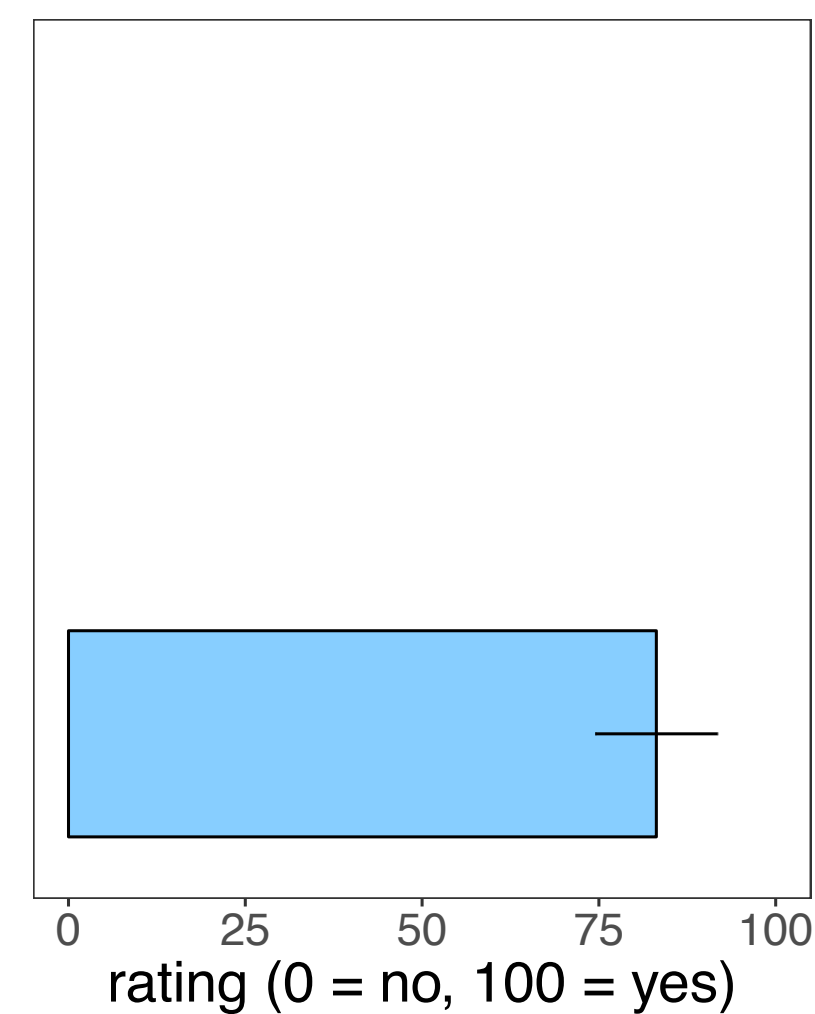


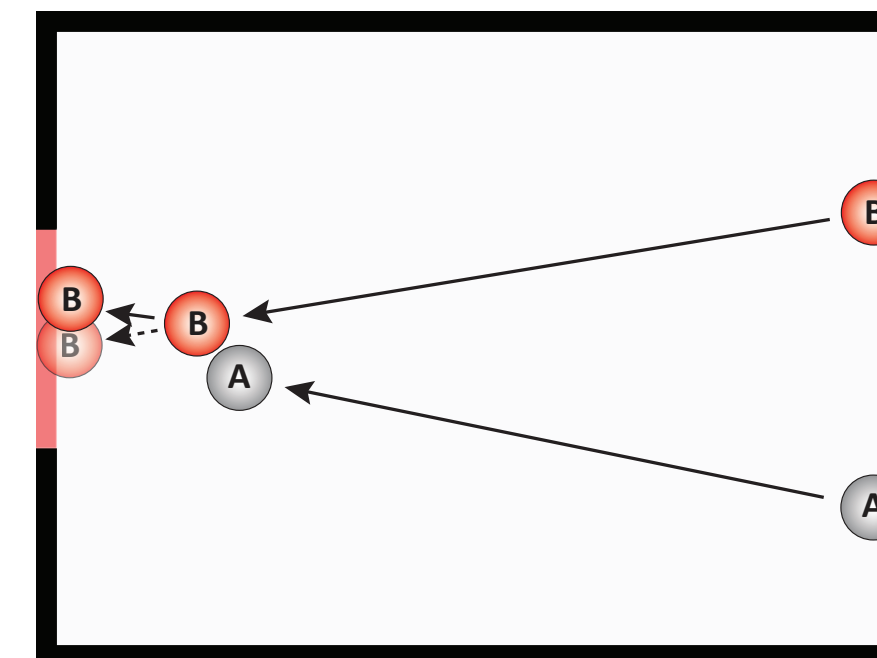
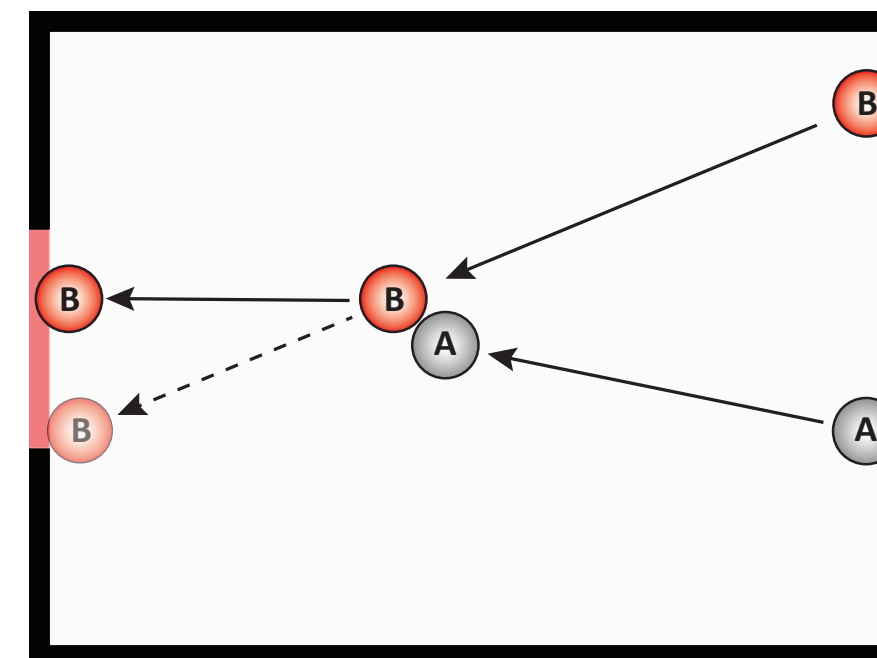
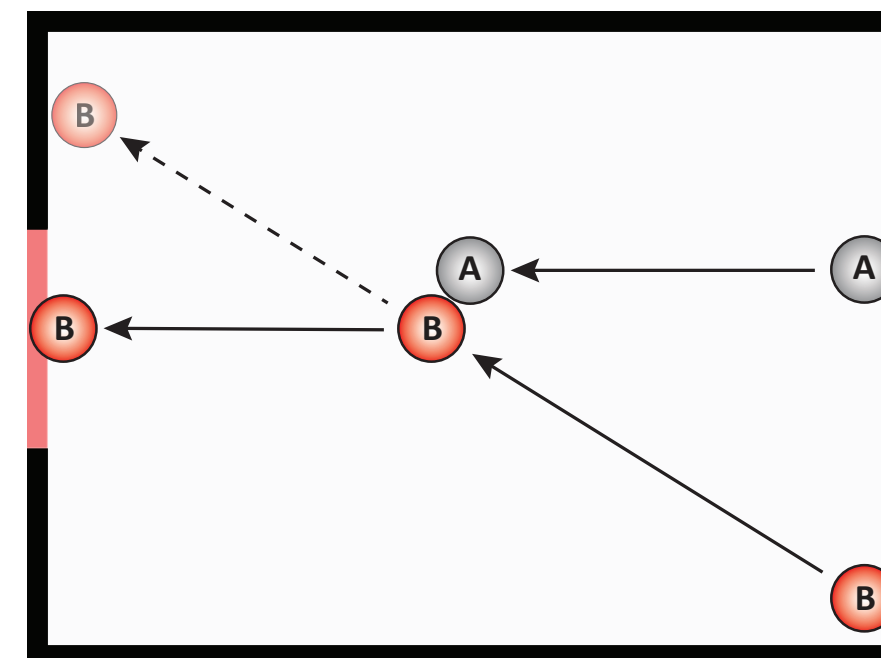


Counterfactual
simulation model

Did A cause B to go
through the gate?

Would B have
missed the gate?

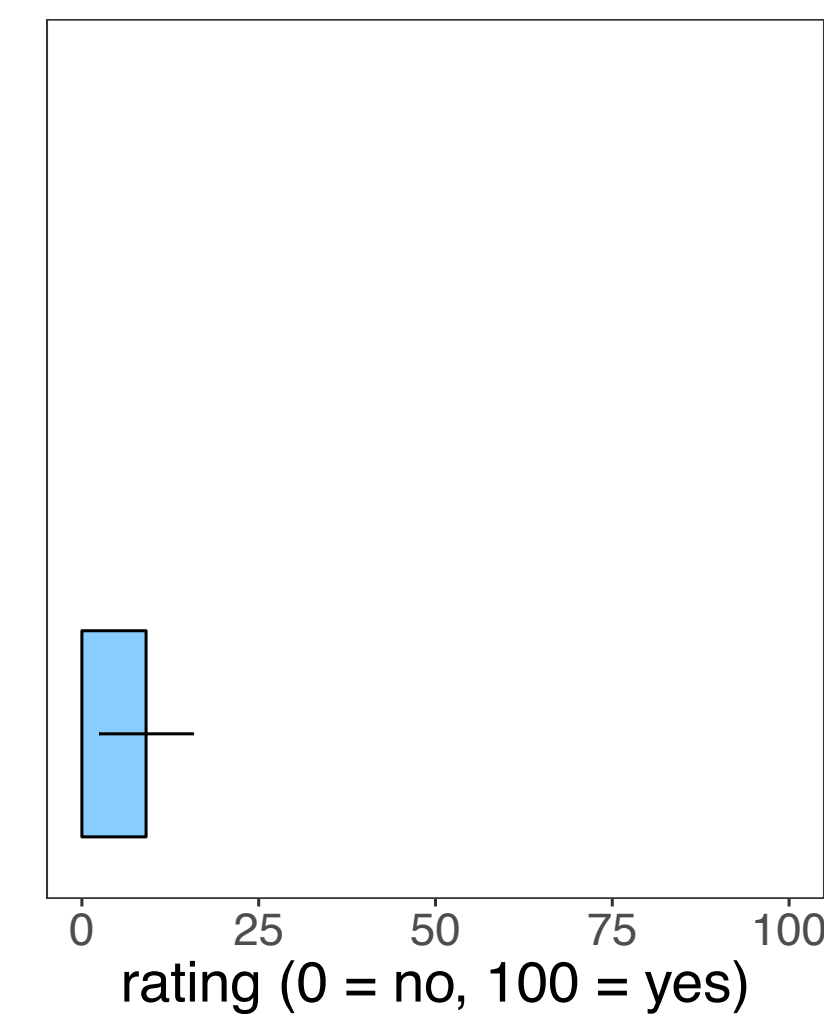
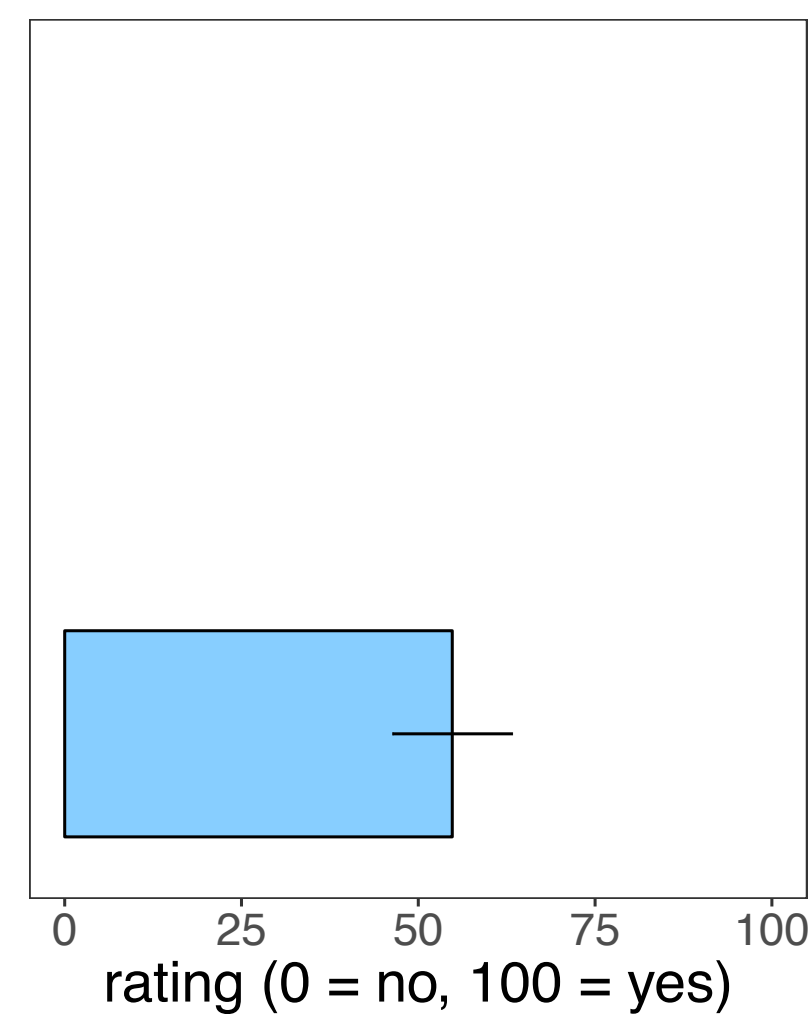
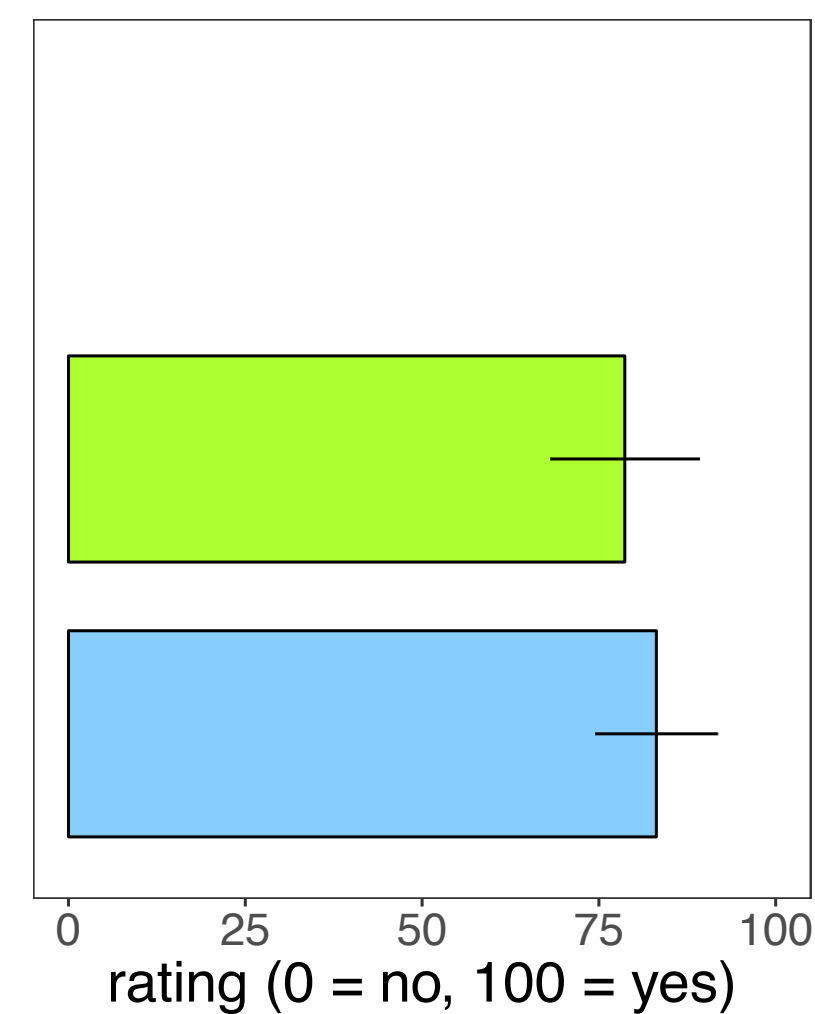


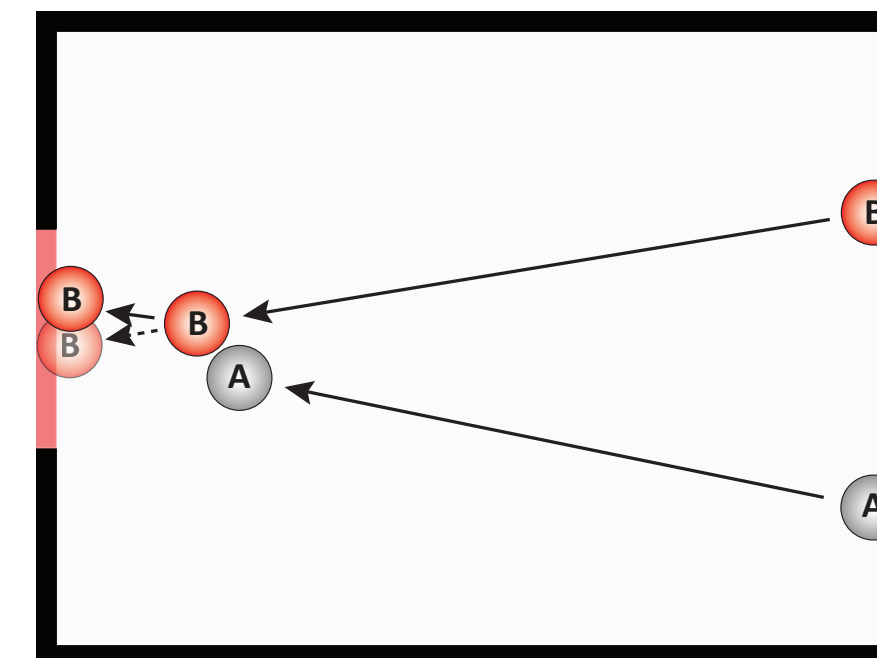
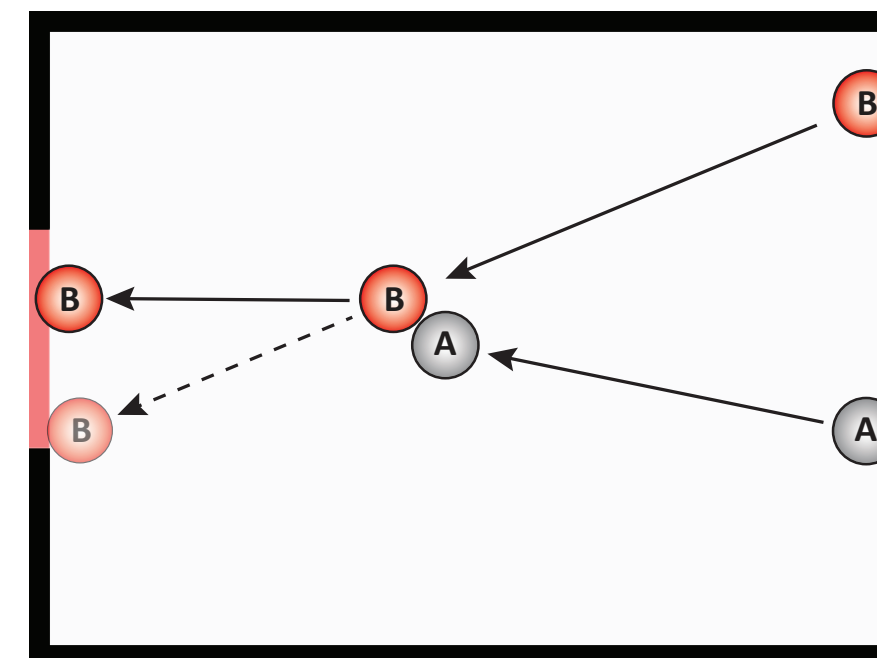
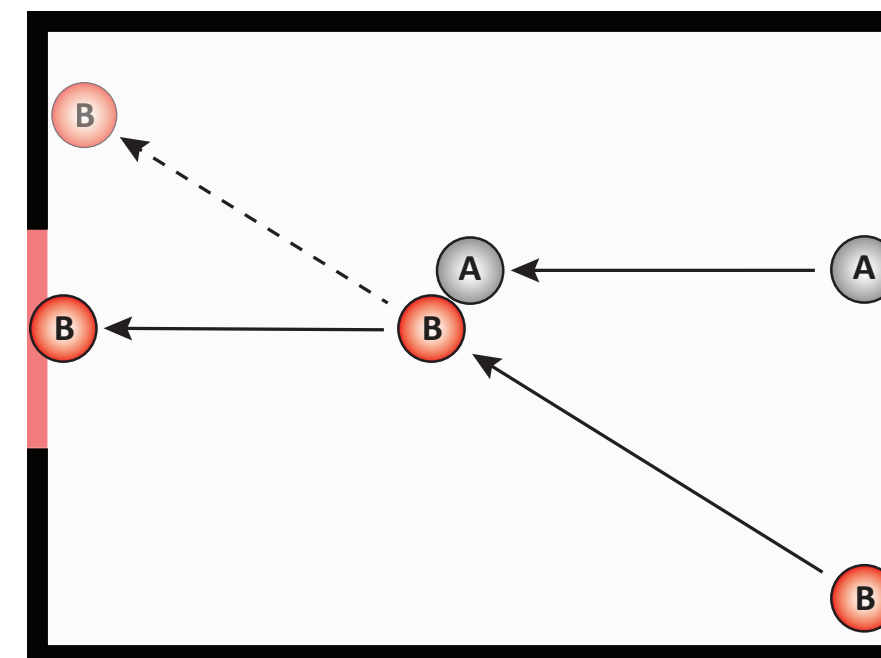


Counterfactual
simulation model

Did A cause B to go
through the gate?

Would B have
missed the gate?

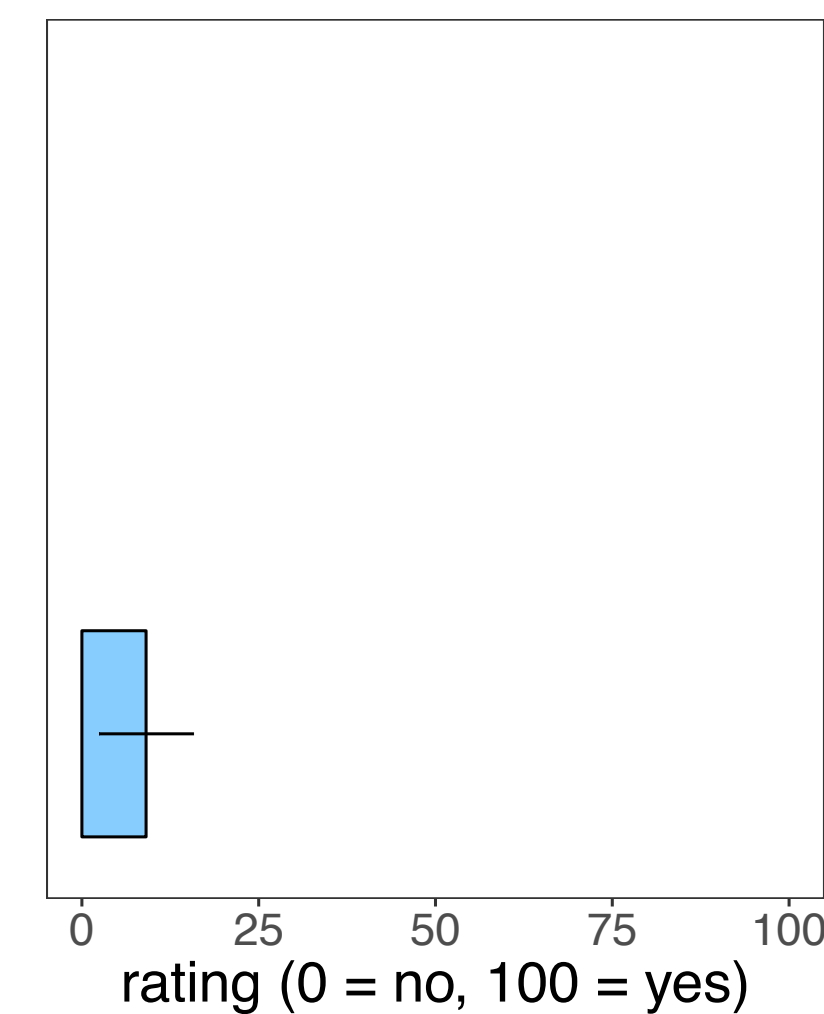
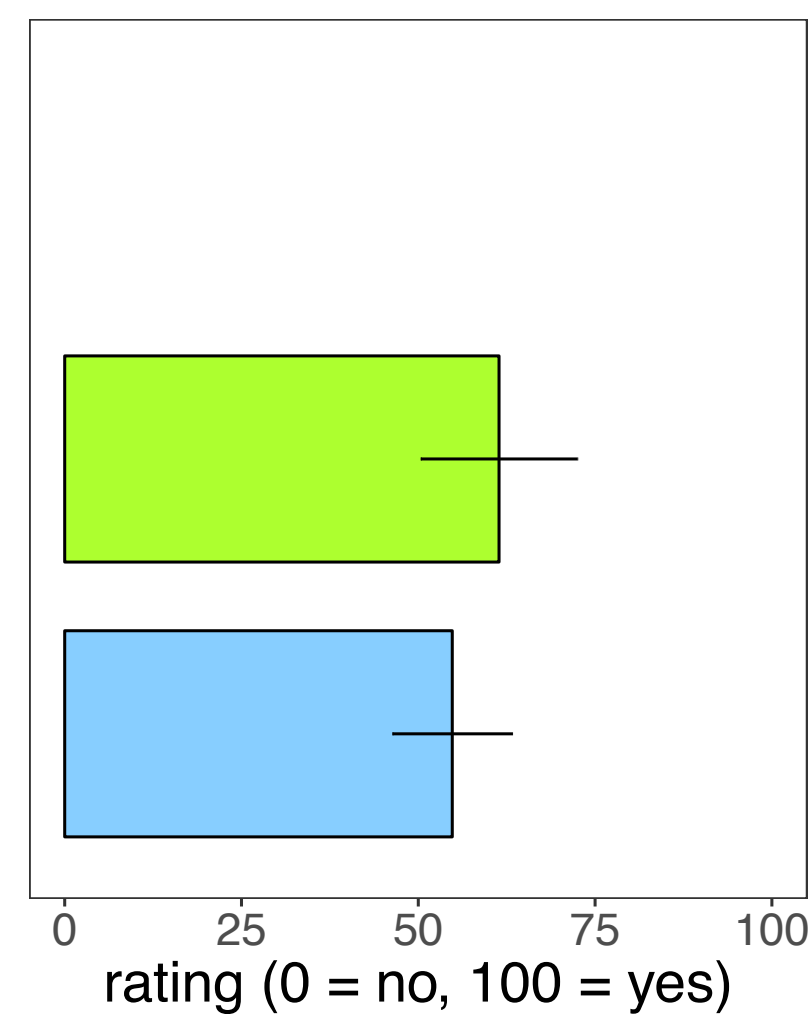
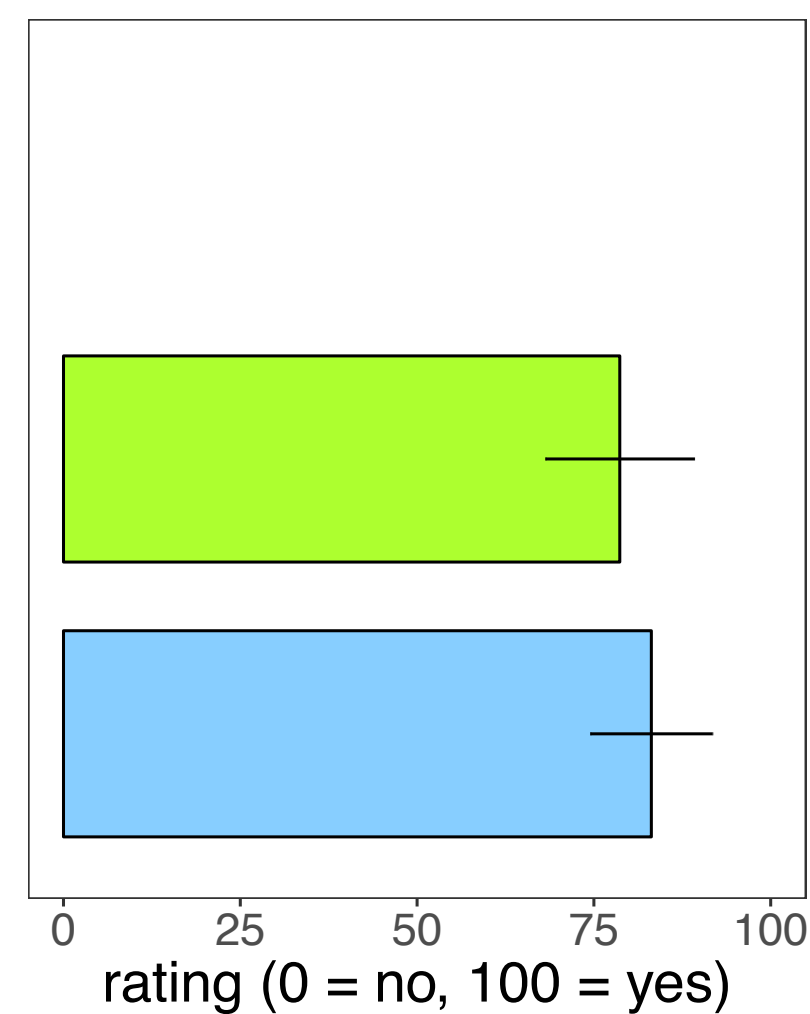


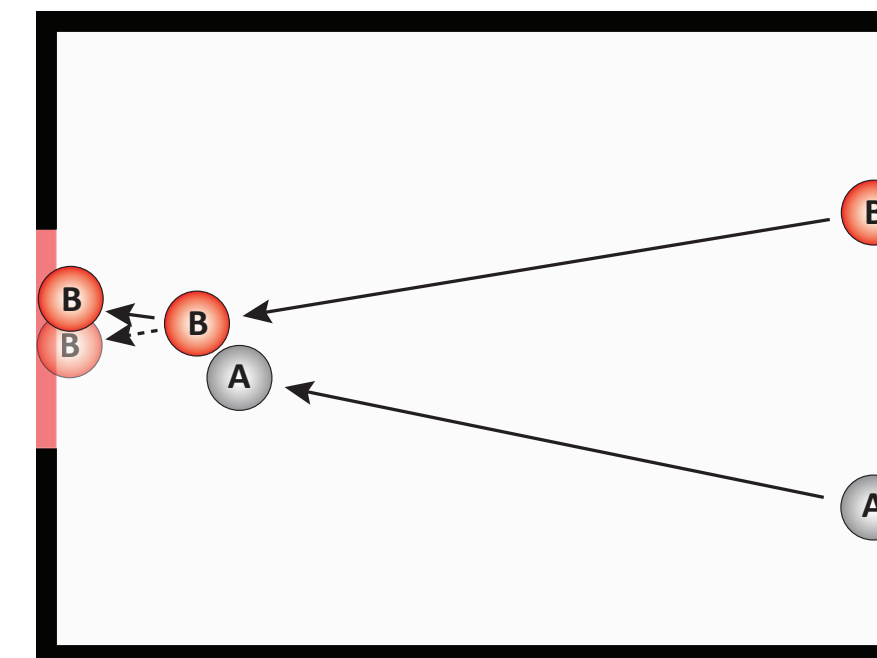
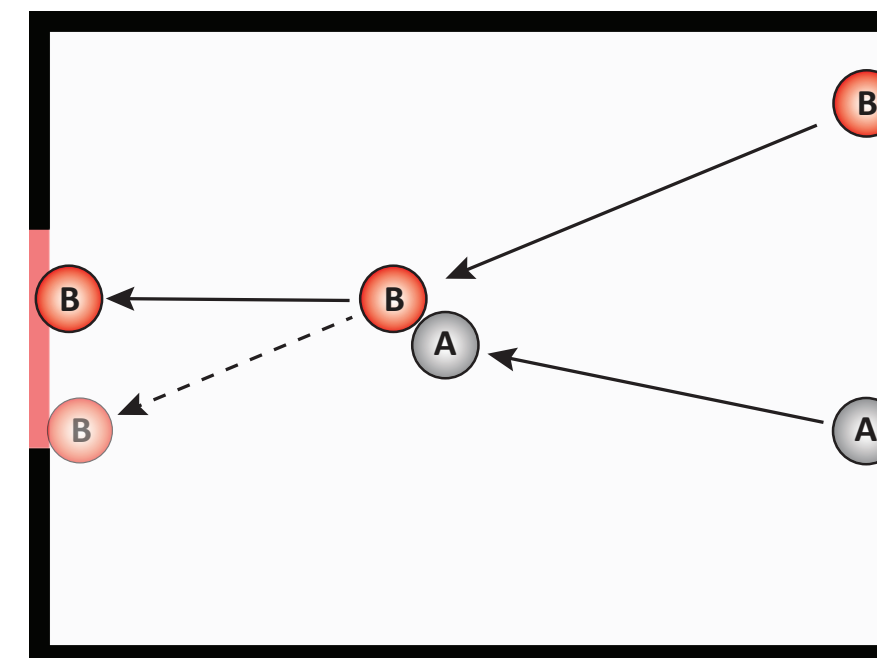
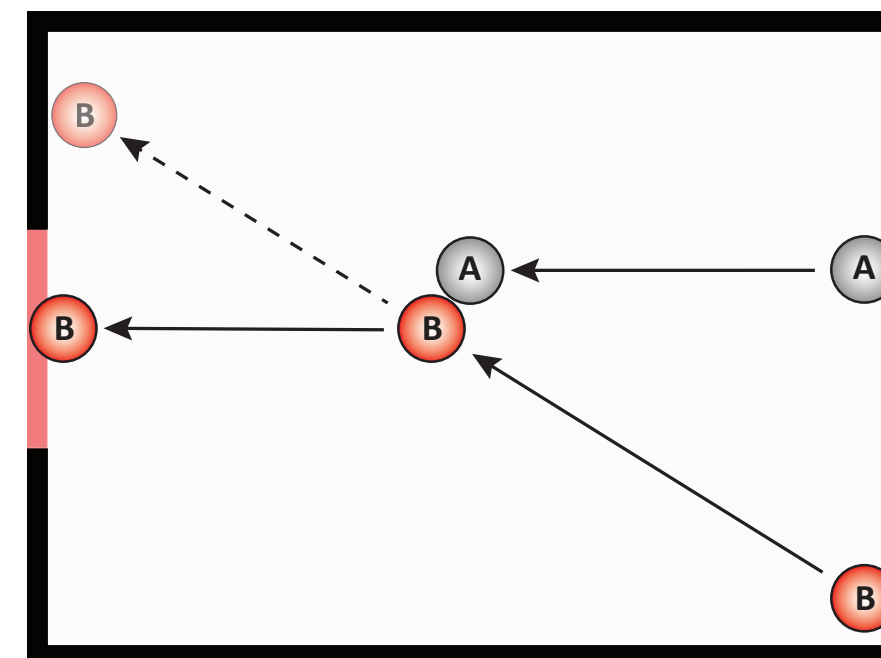


Counterfactual
simulation model

Did A cause B to go
through the gate?

Would B have
missed the gate?

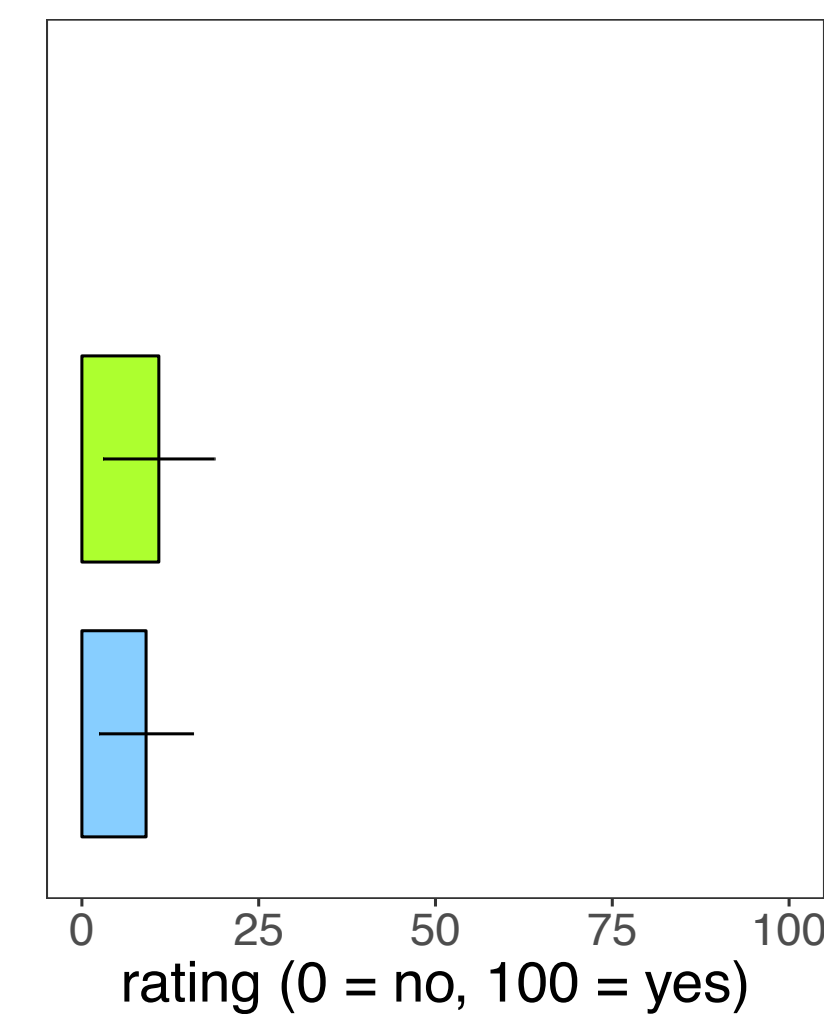
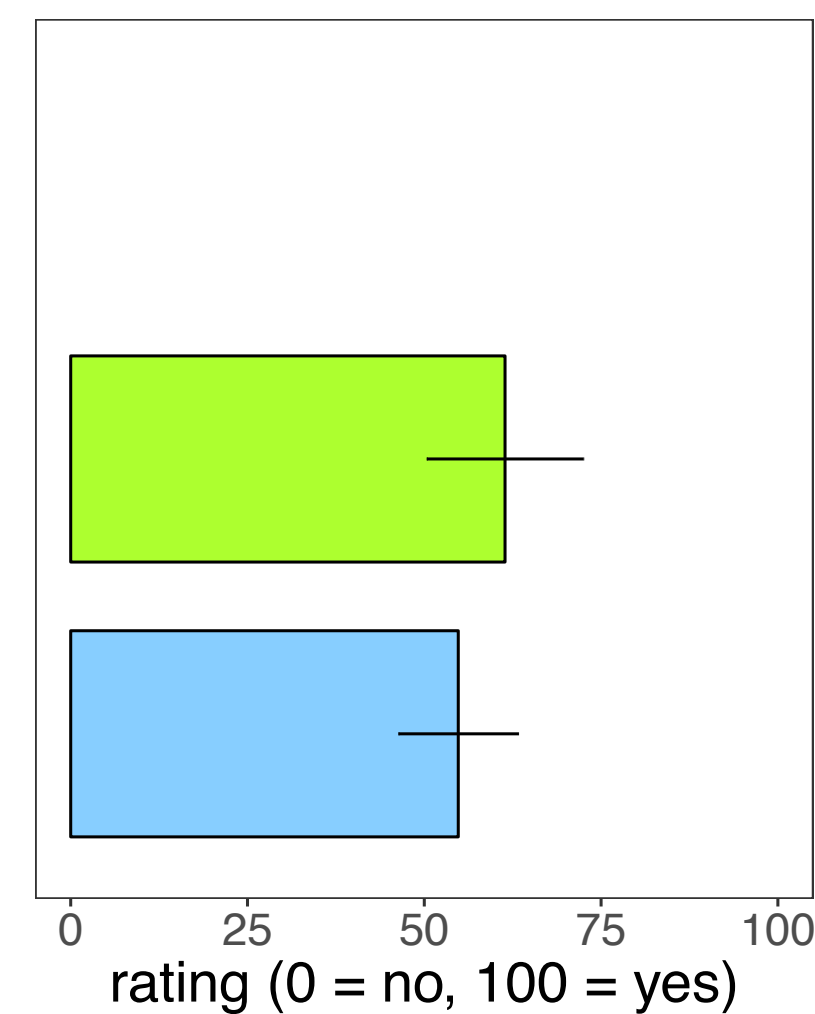
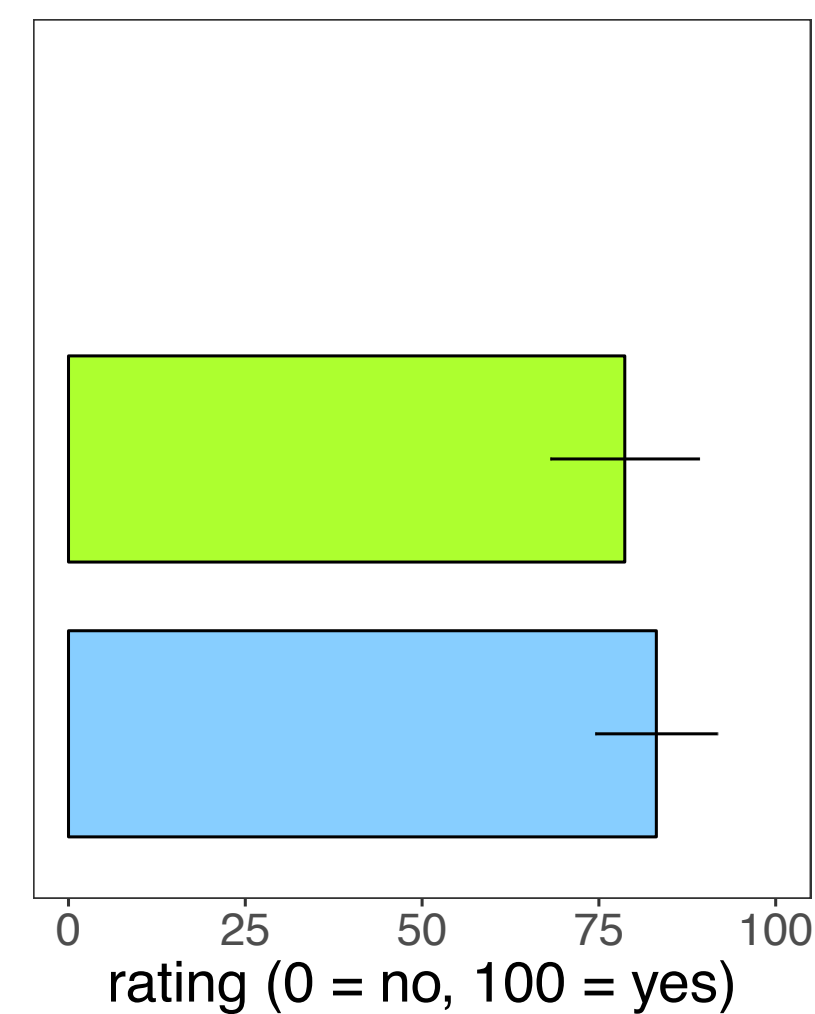


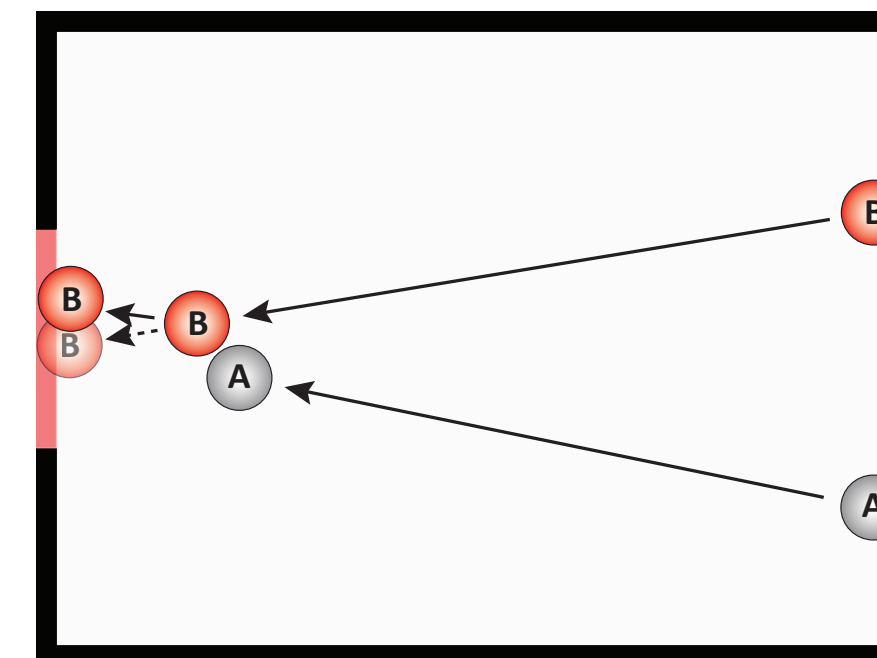
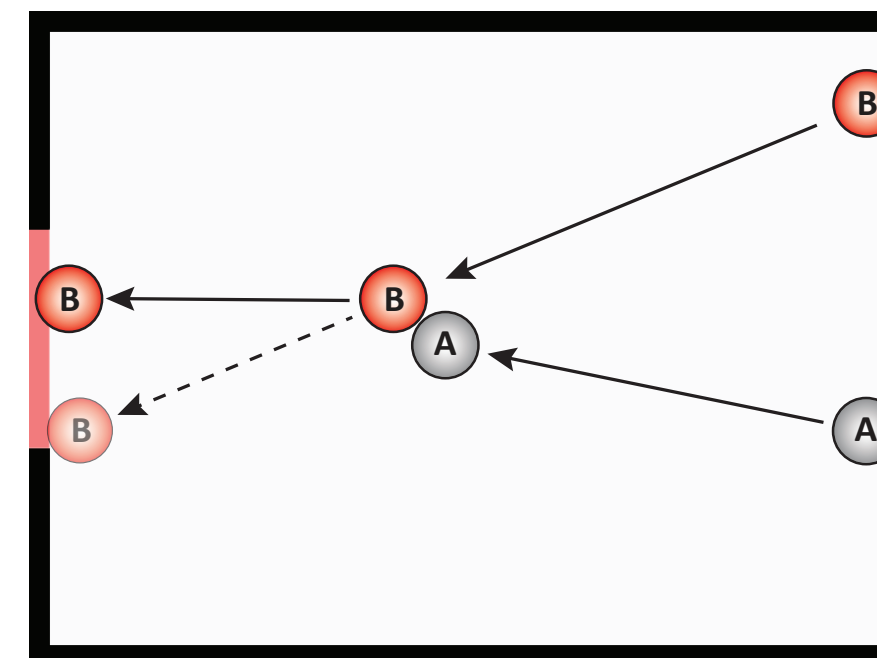
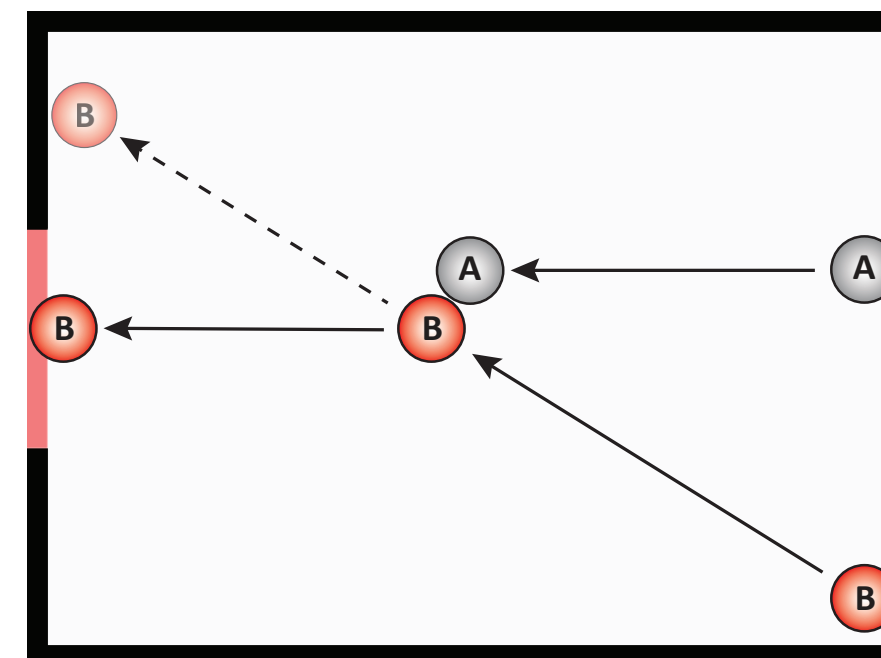


Counterfactual
simulation model

Did A cause B to go
through the gate?

Would B have
missed the gate?

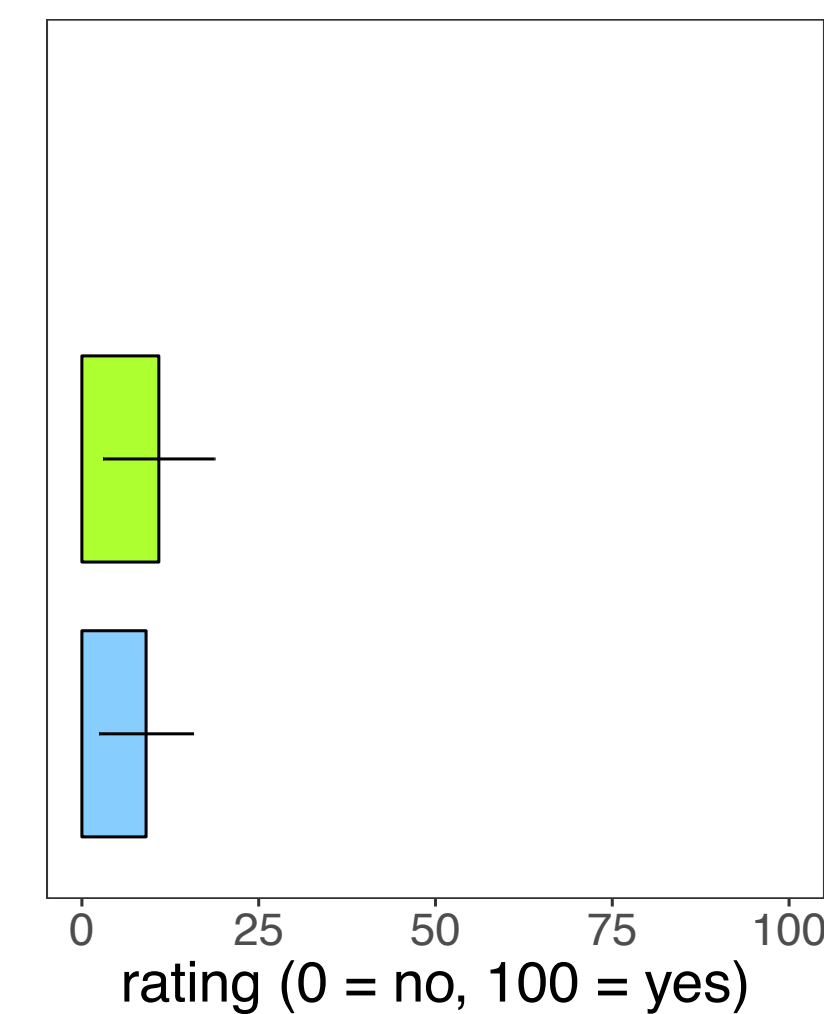
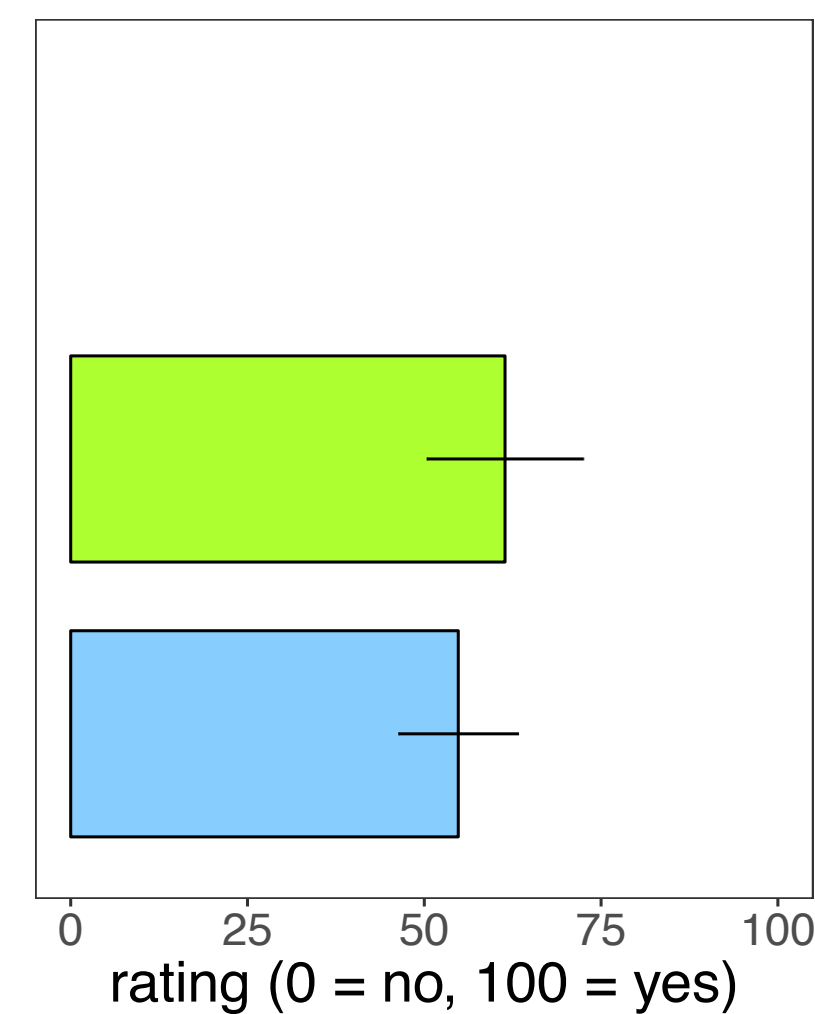
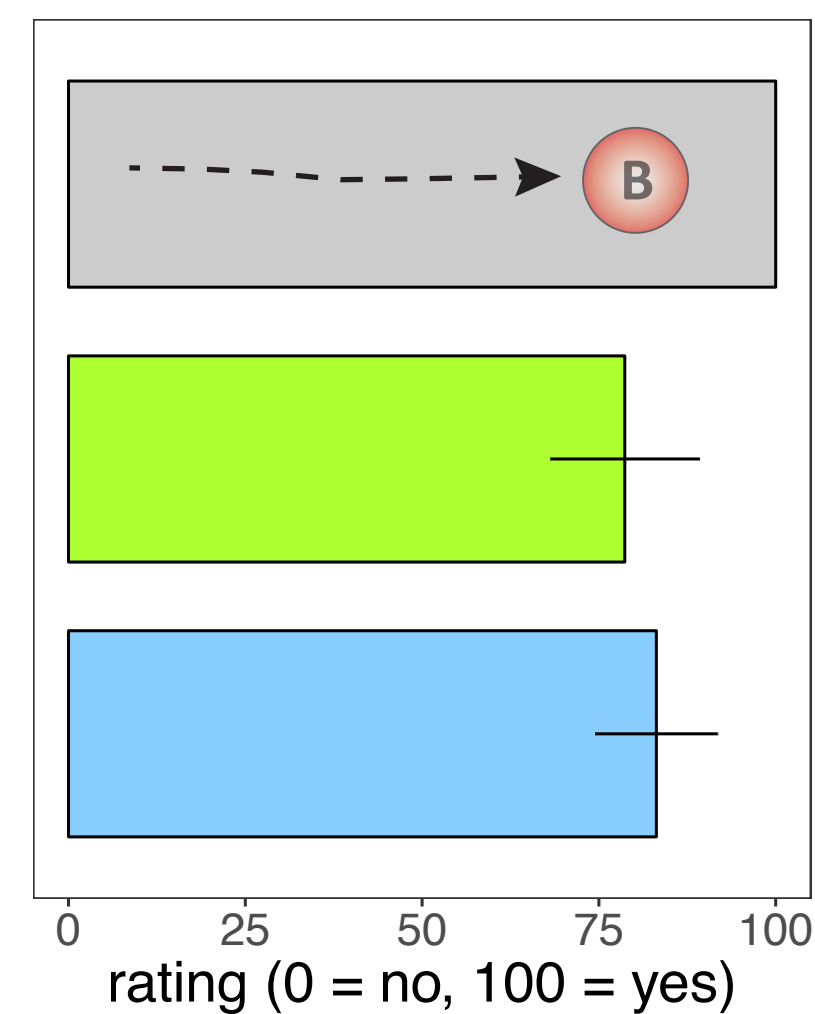


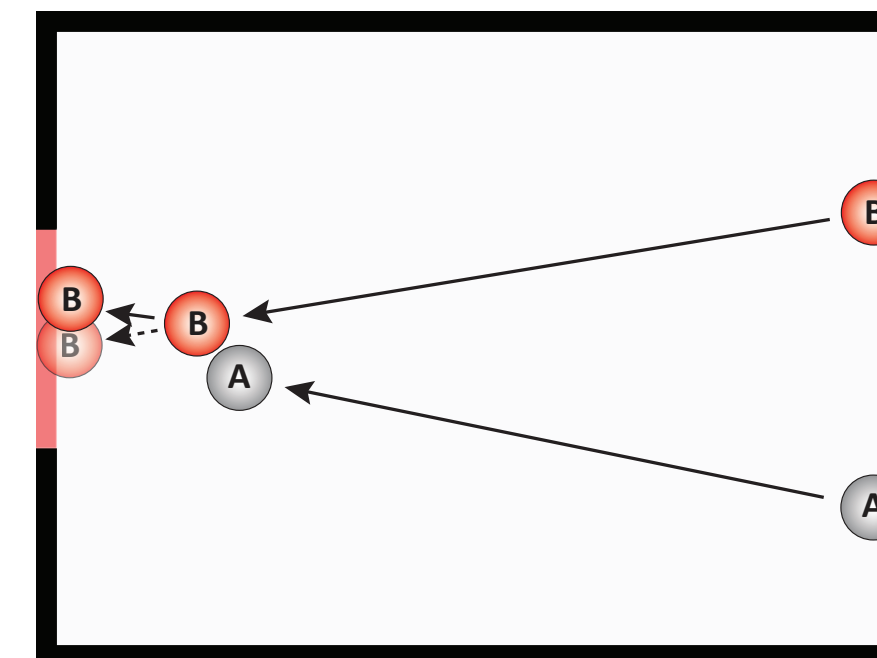
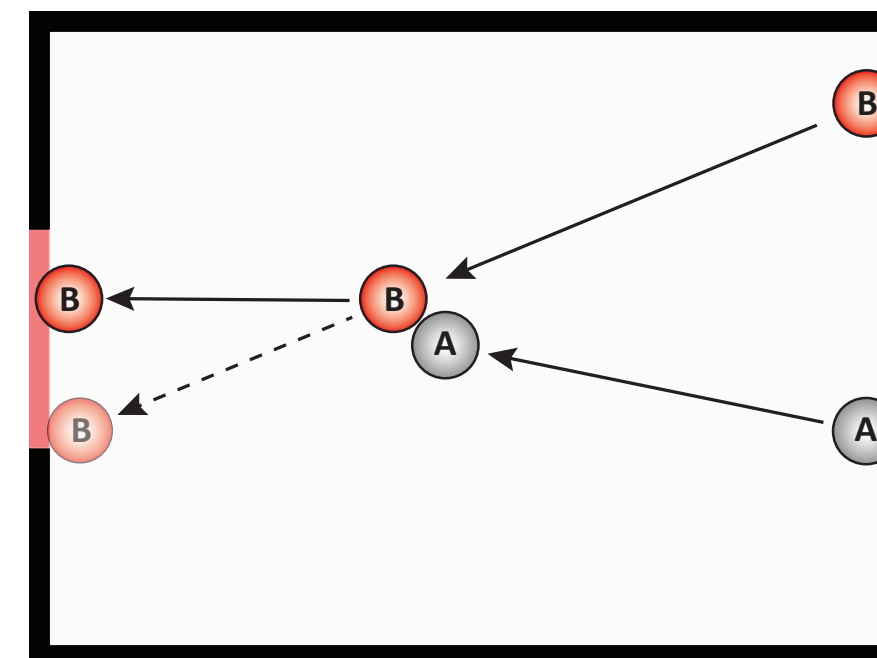
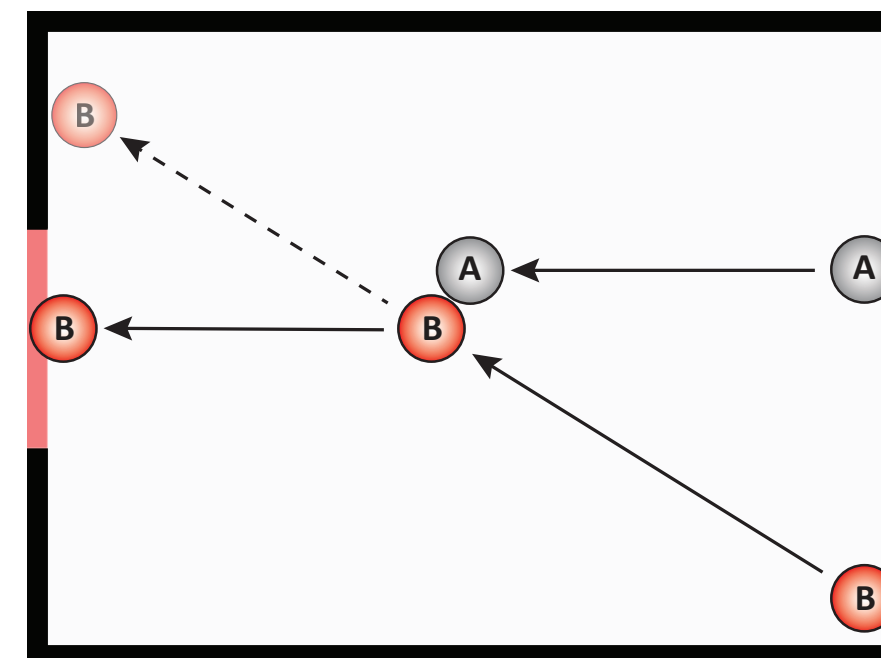


Counterfactual
simulation model

Did A cause B to go
through the gate?

Would B have
missed the gate?

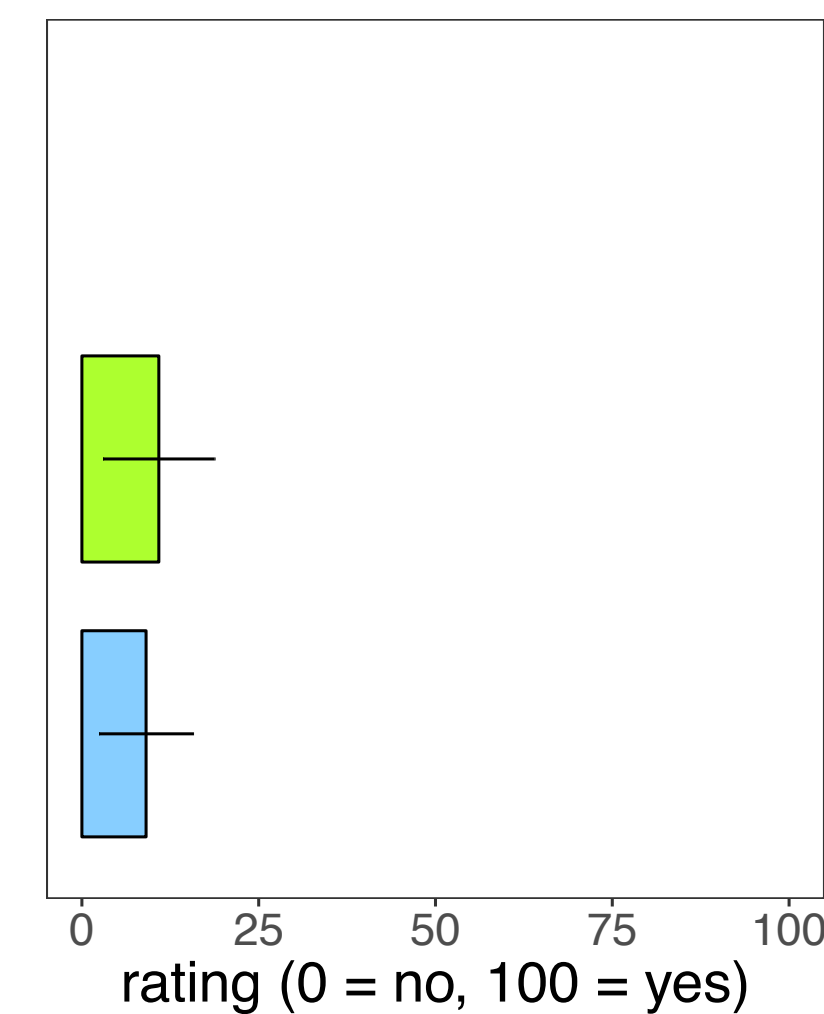
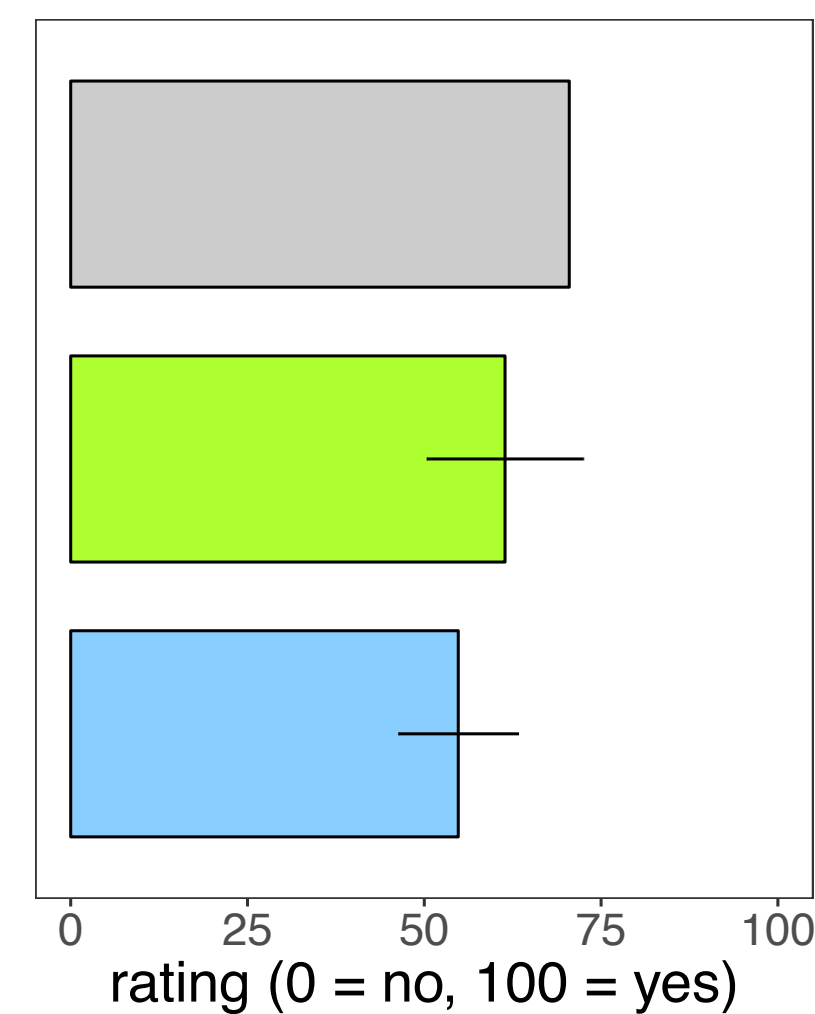
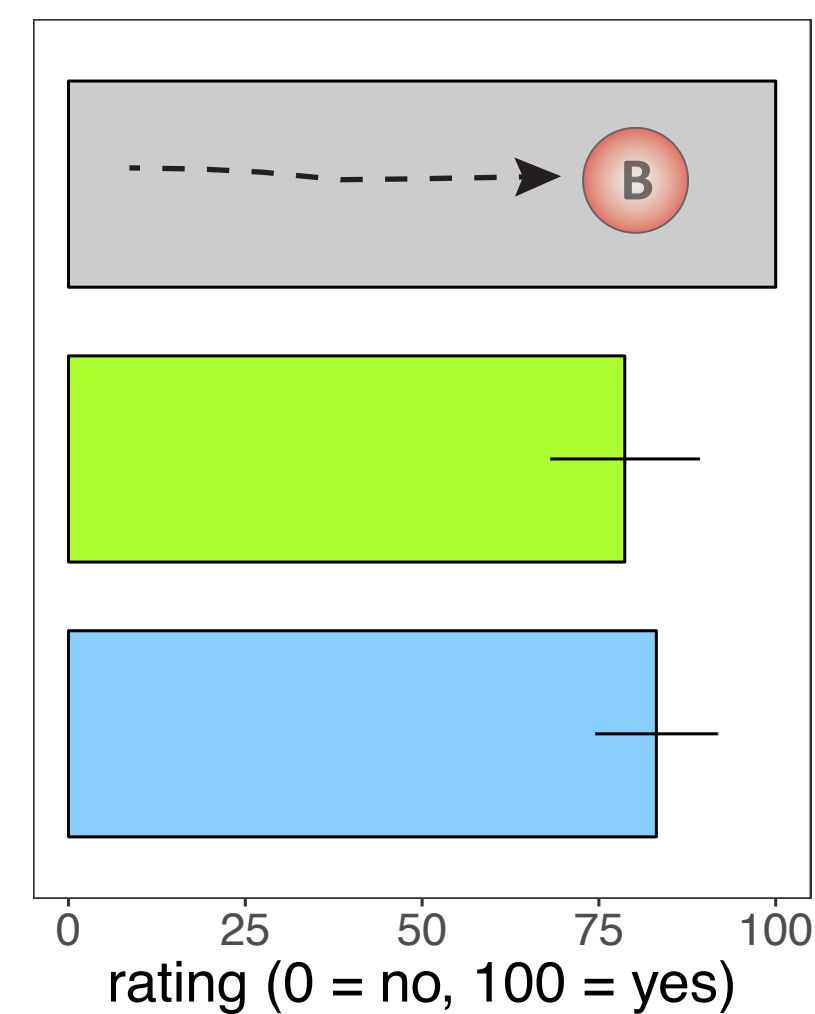


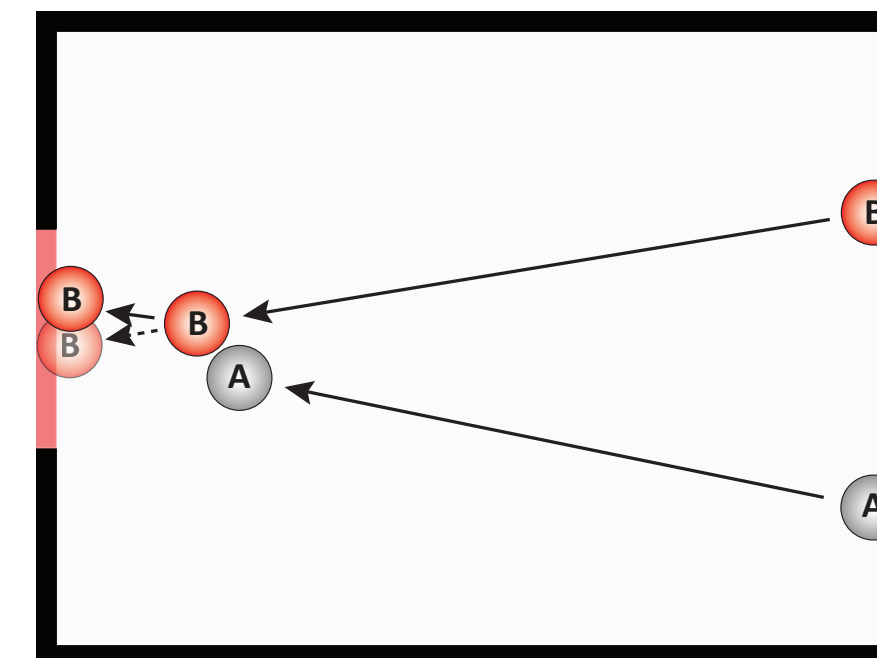
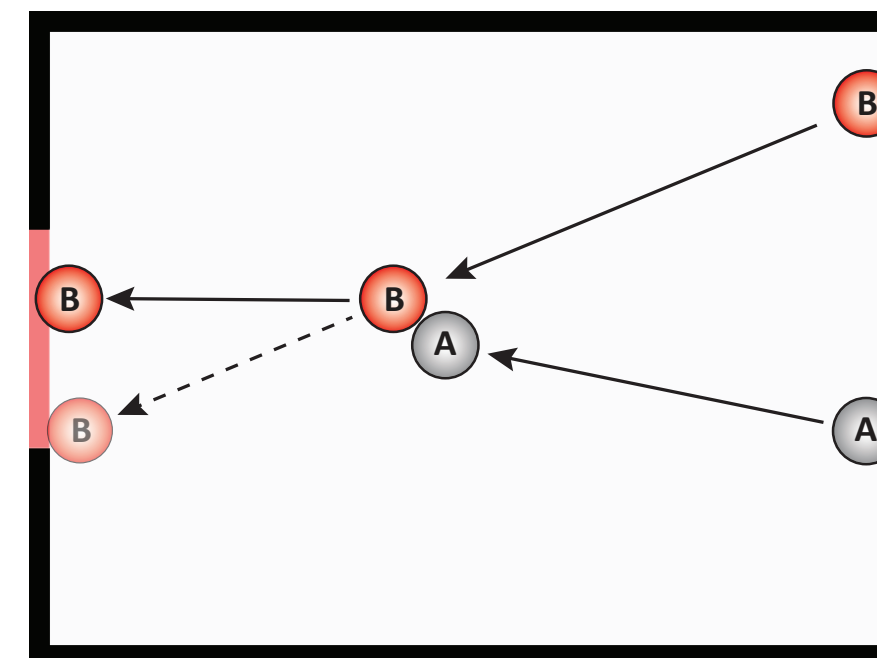
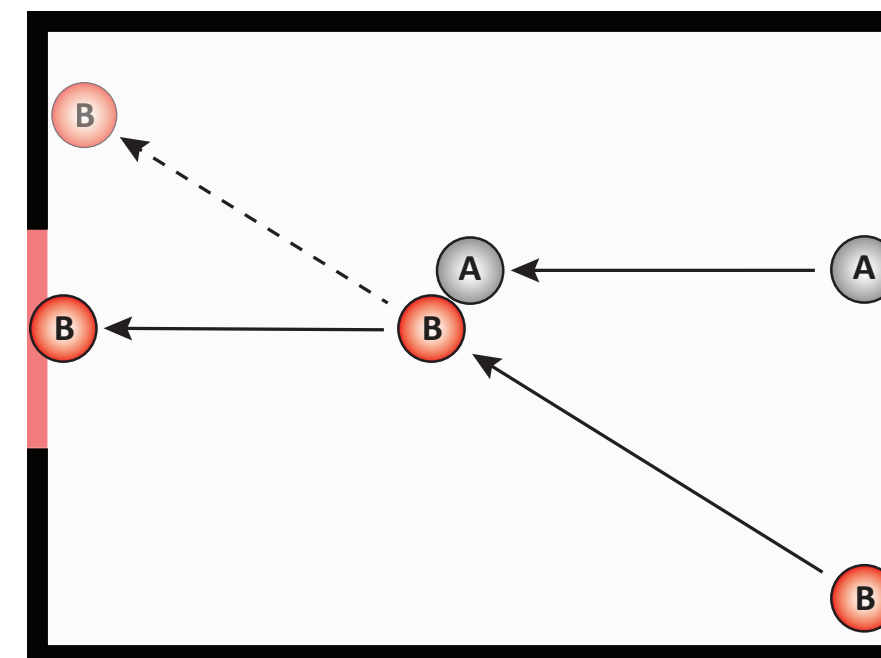


**Counterfactual
simulation model**

**Did A cause B to go
through the gate?**

**Would B have
missed the gate?**

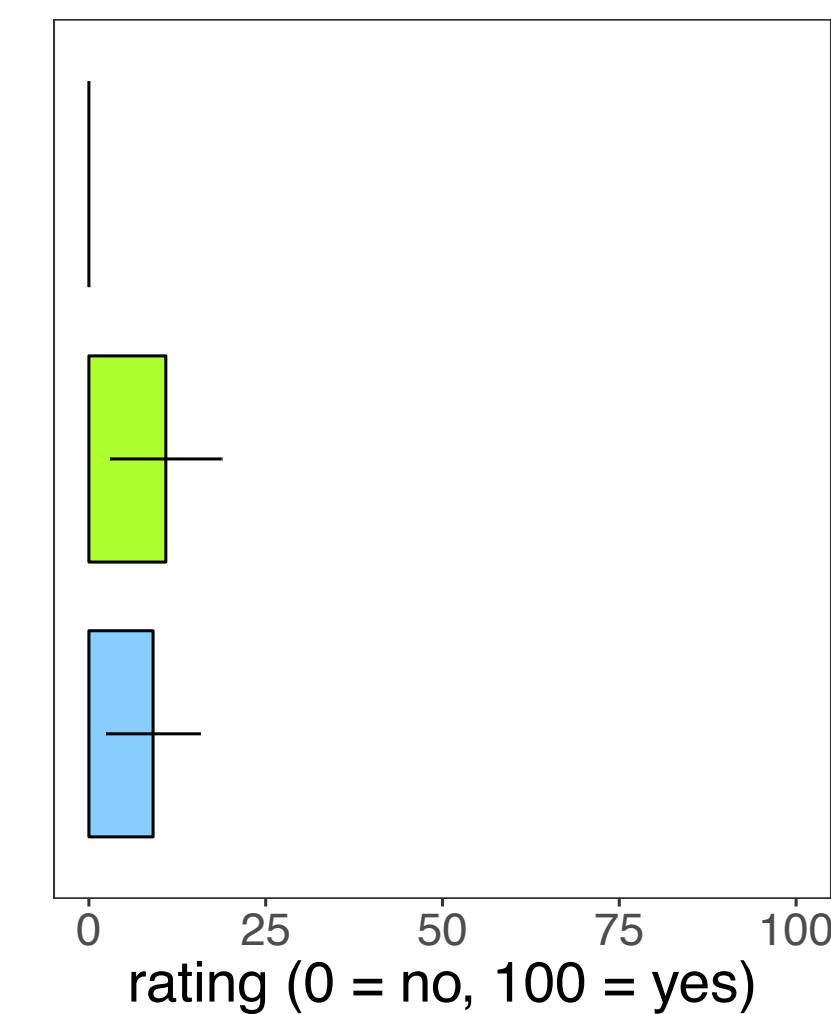
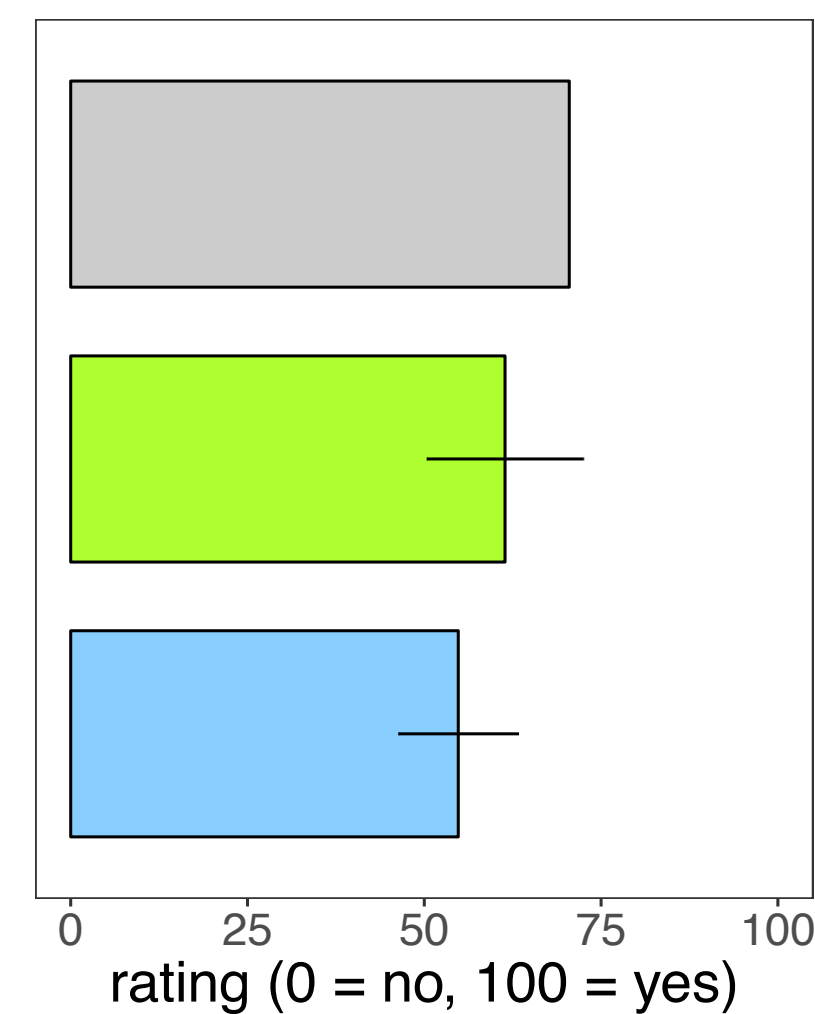
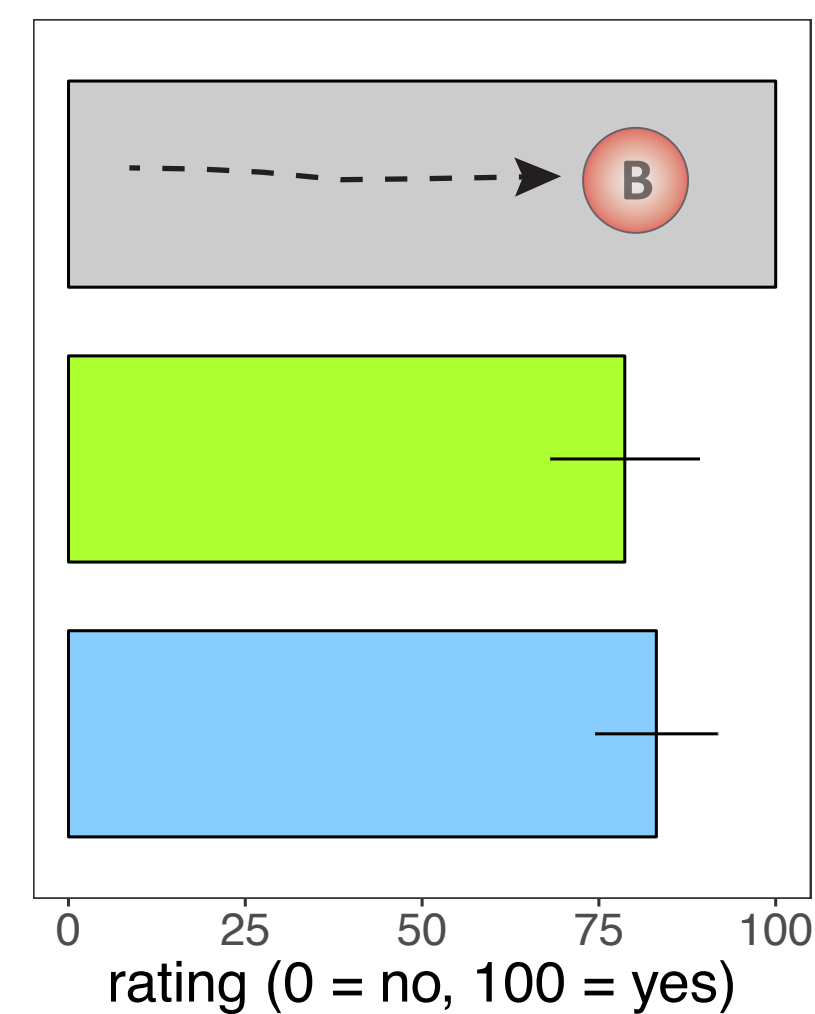


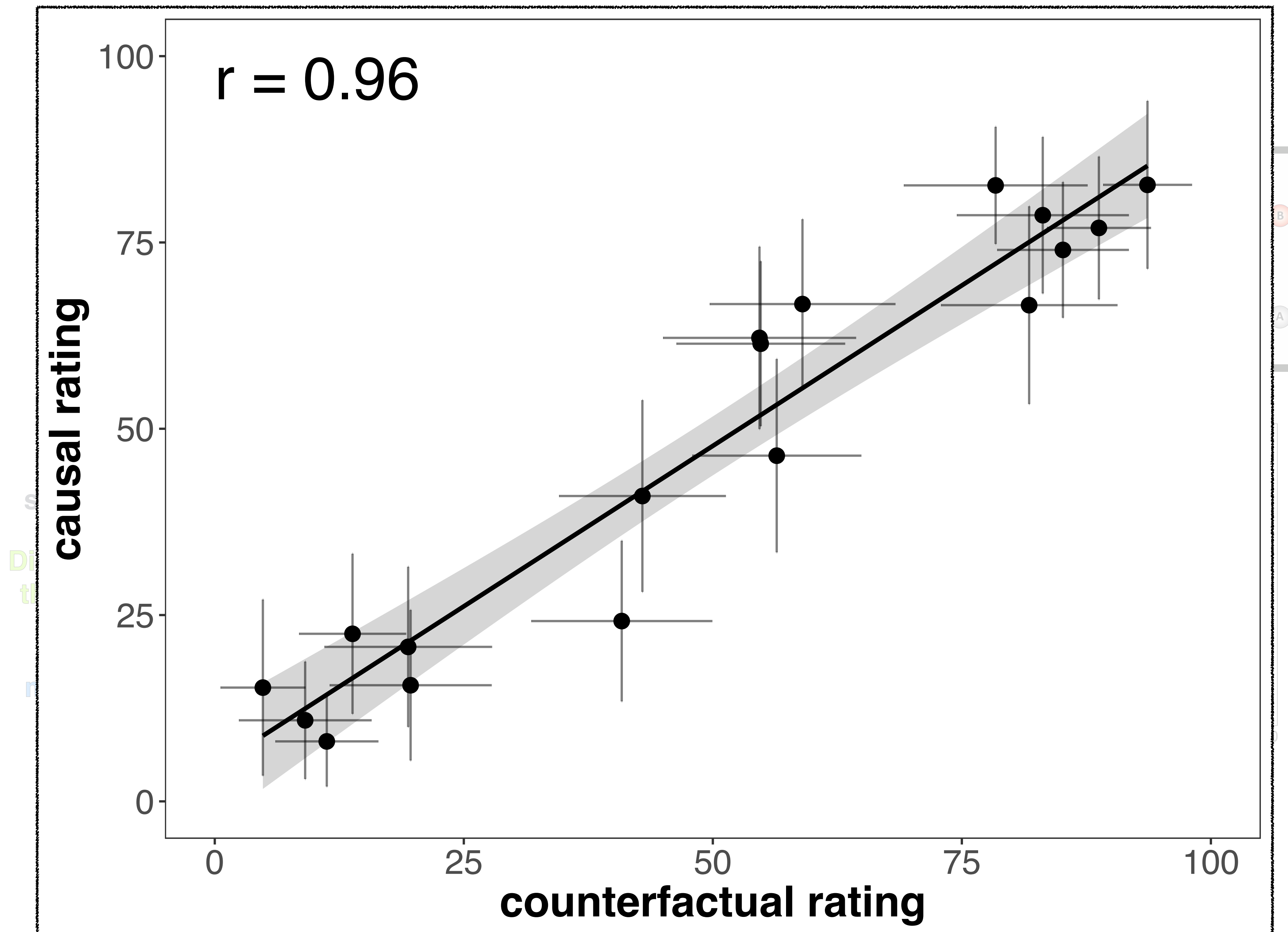


Counterfactual
simulation model

Did A cause B to go
through the gate?

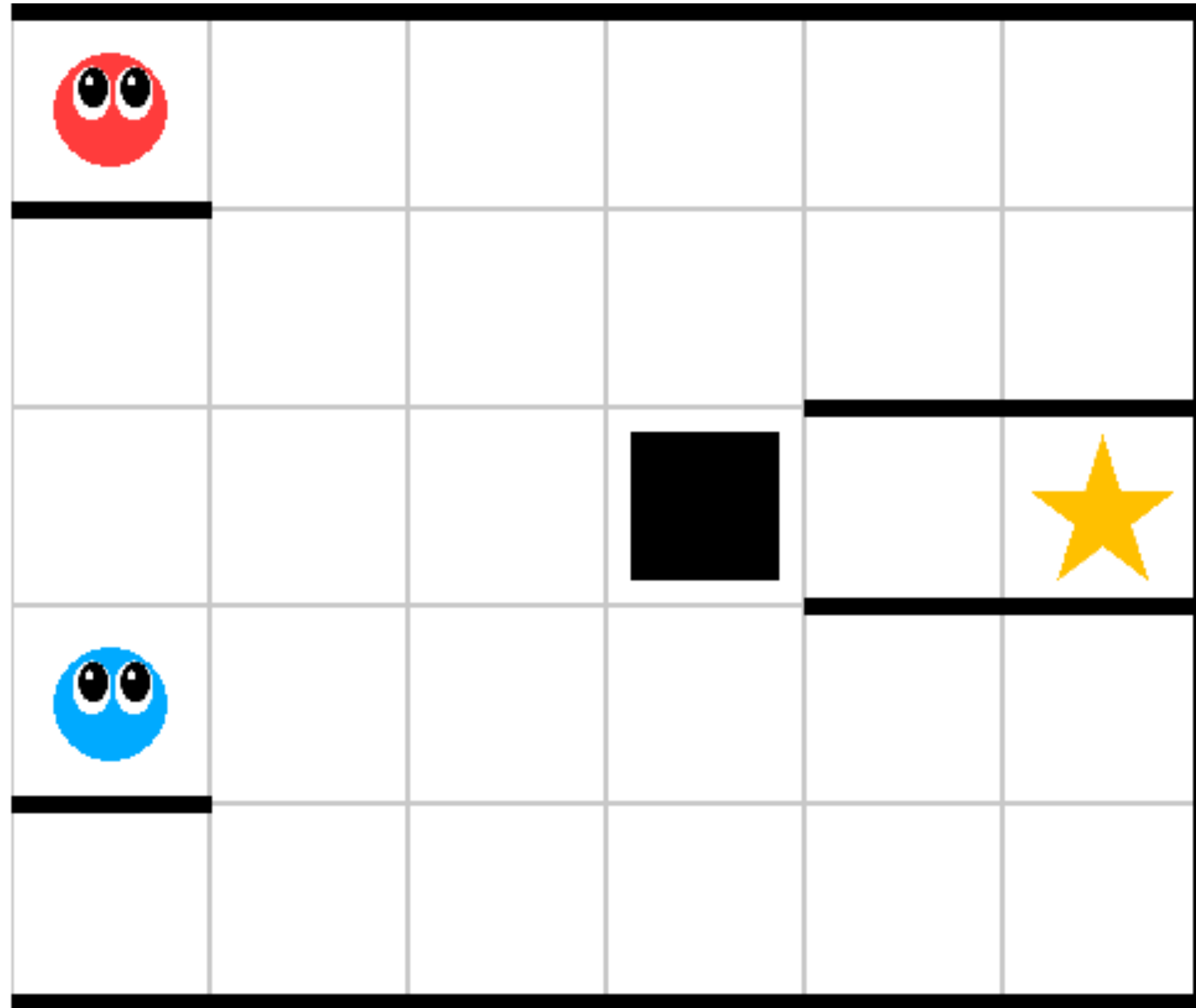
Would B have
missed the gate?





Deep dive: Counterfactual simulation for responsibility judgments

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

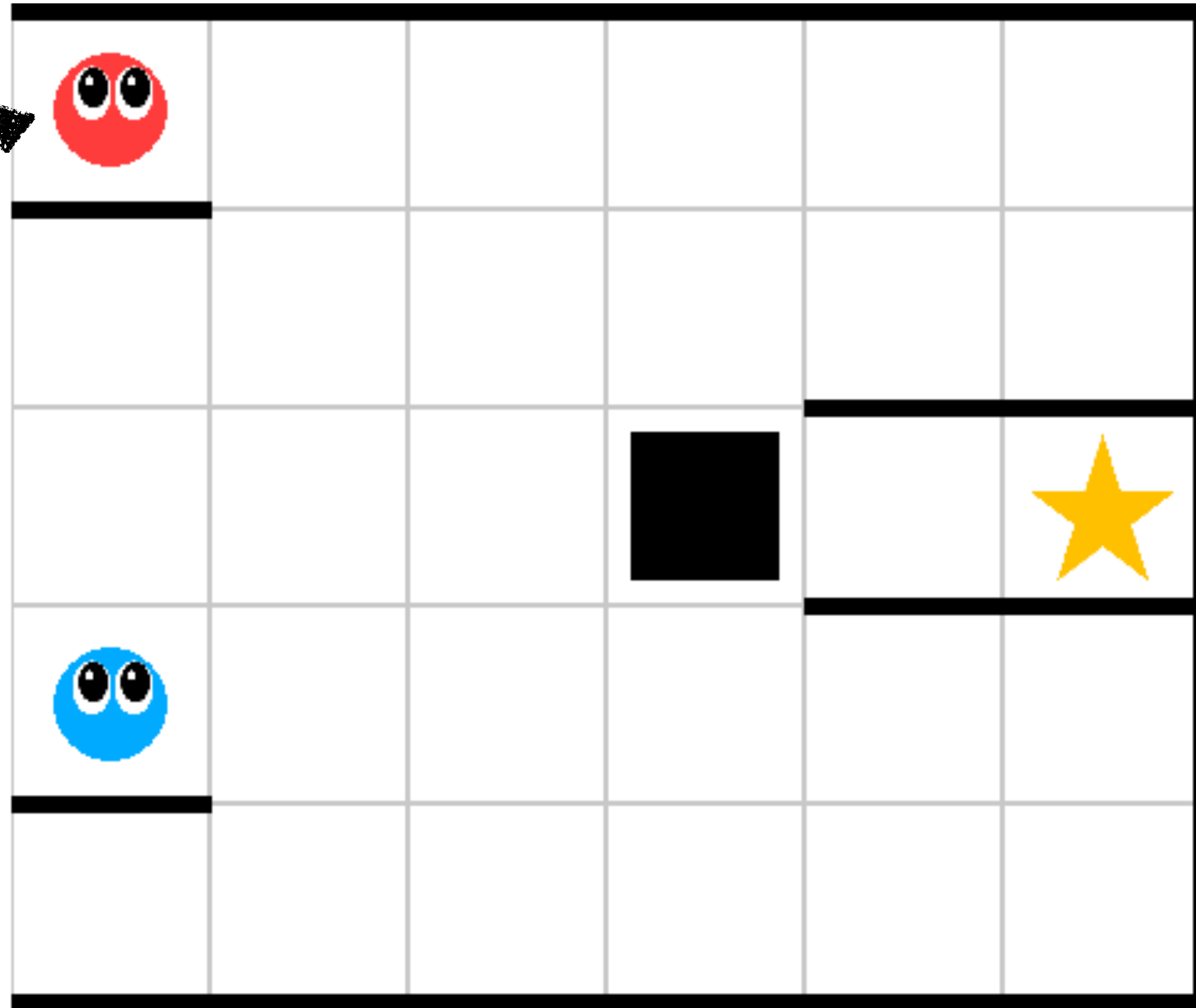


time left:

10

result:

wants to get
to the star

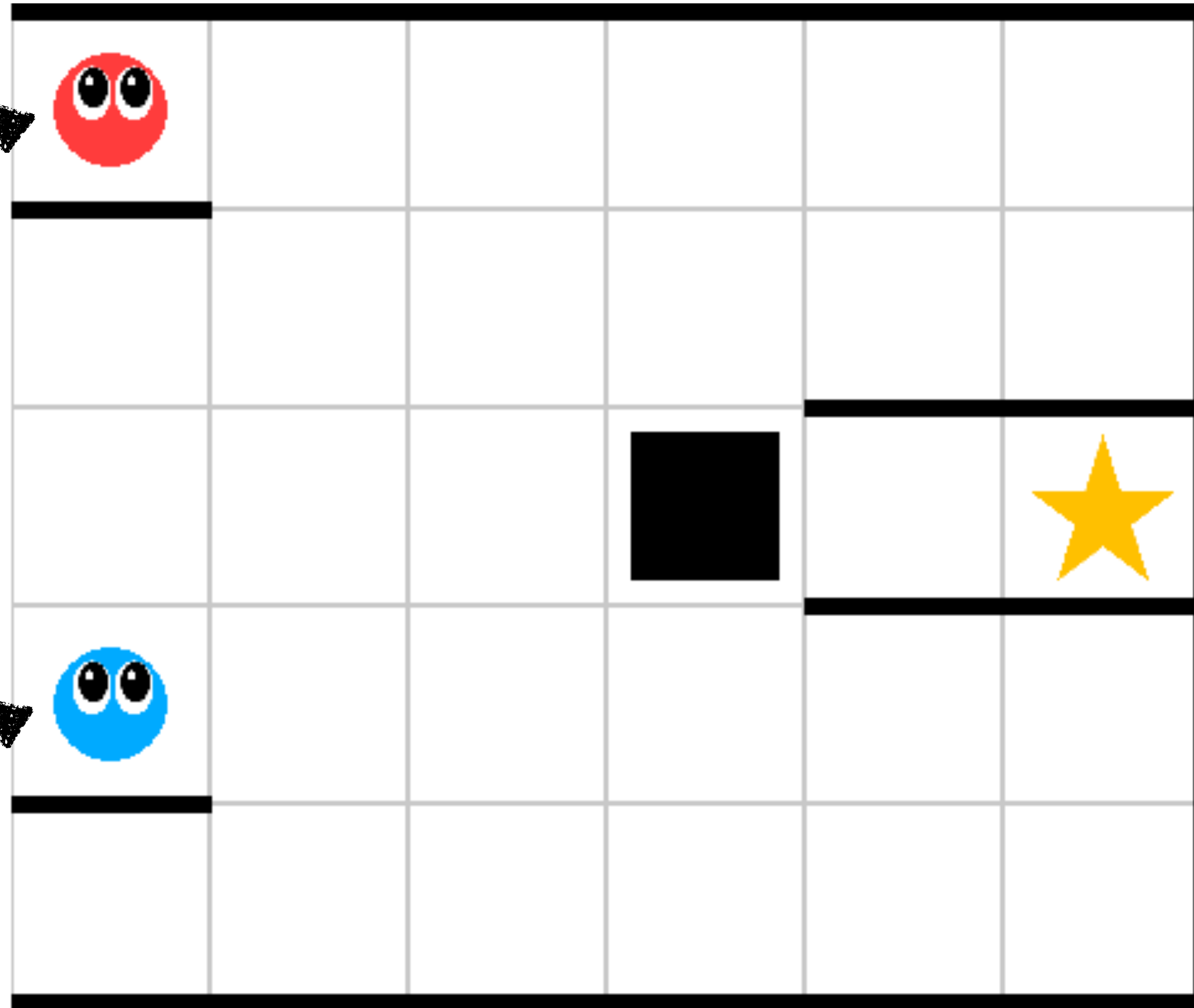


time left:

10

result:

wants to get
to the star



time left:

10

result:

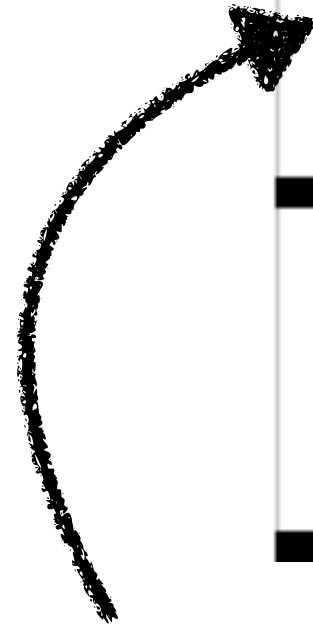
wants to help or
hinder **red**



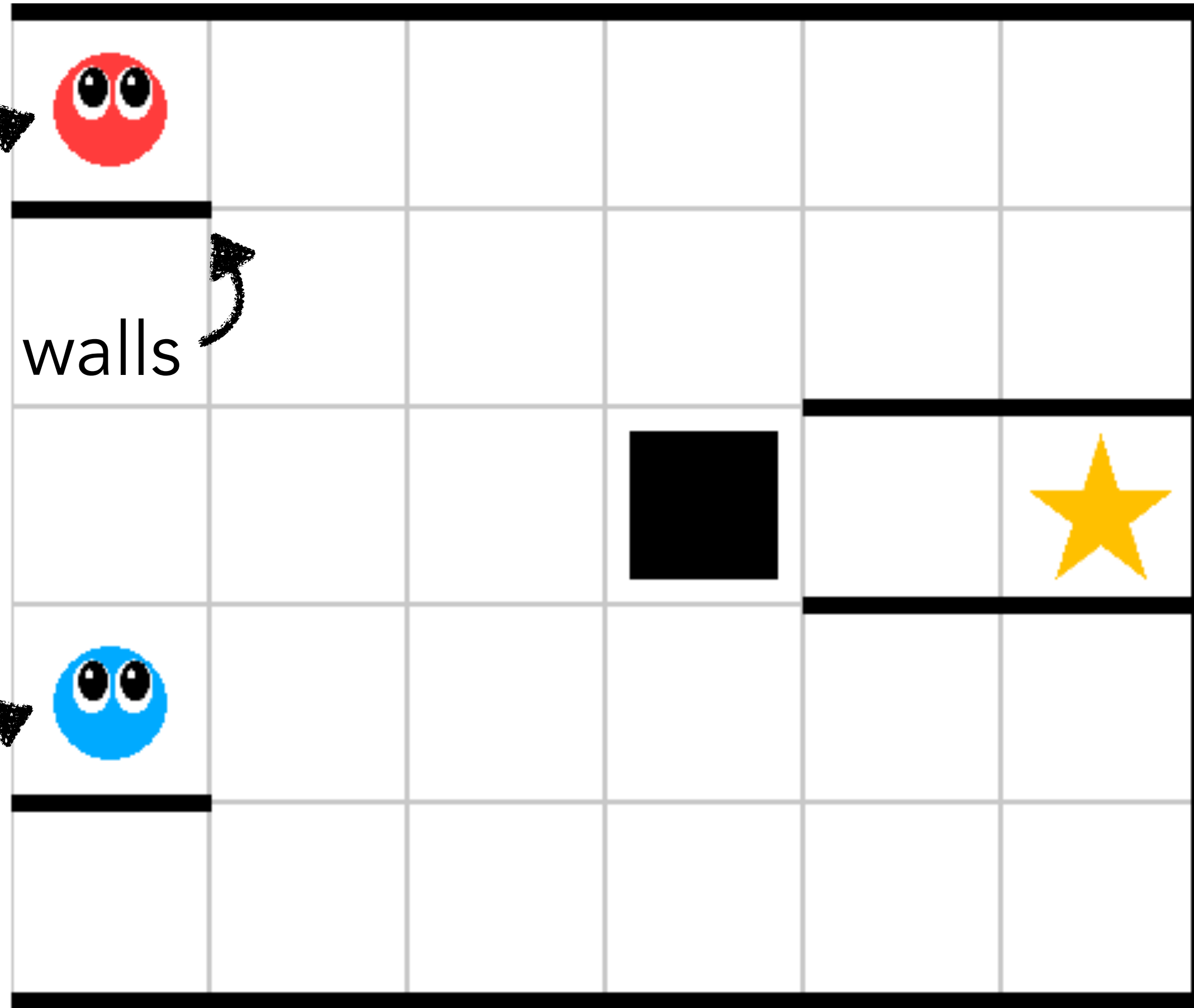
wants to get
to the star



static walls



wants to help or
hinder **red**

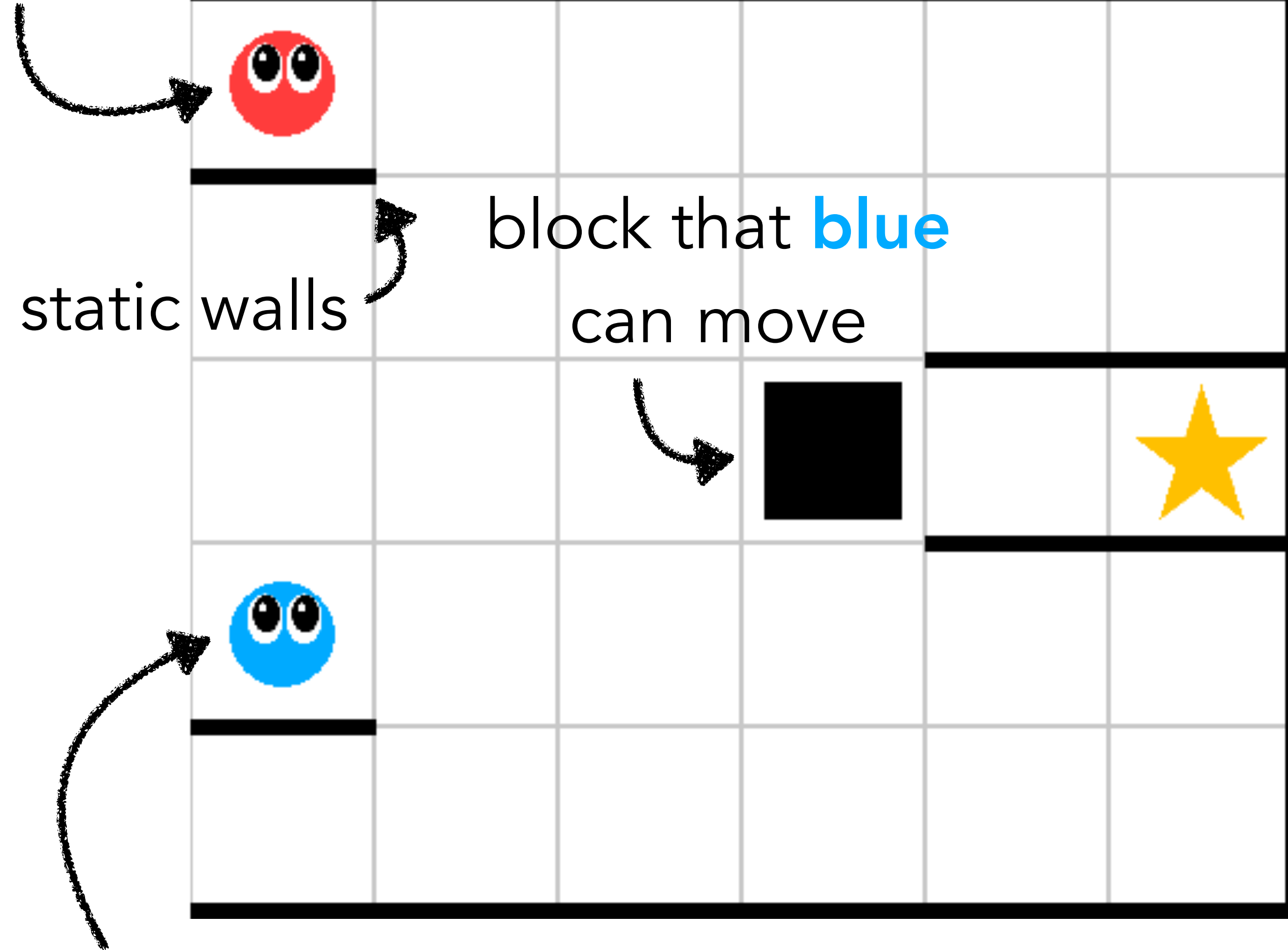


time left:

10

result:

wants to get
to the star



static walls

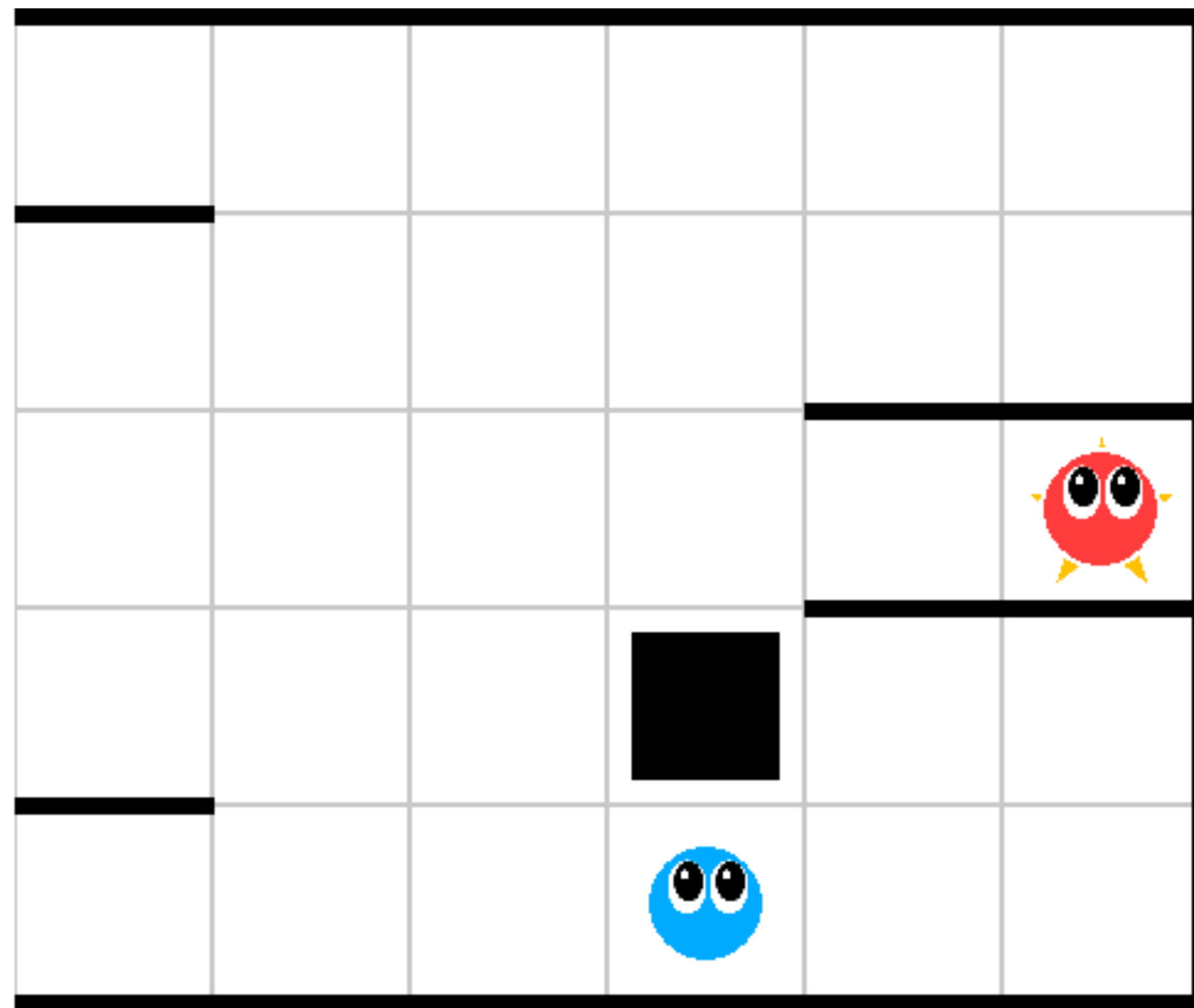
block that **blue**
can move

time left:
10

result:

wants to help or
hinder **red**

Watch Clip 2



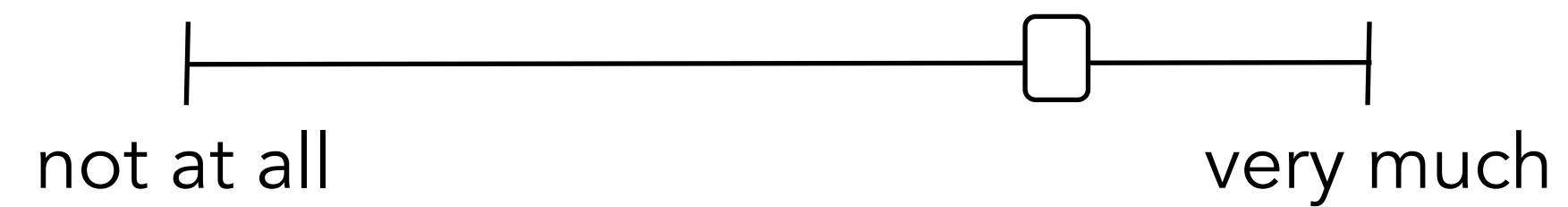
time left:

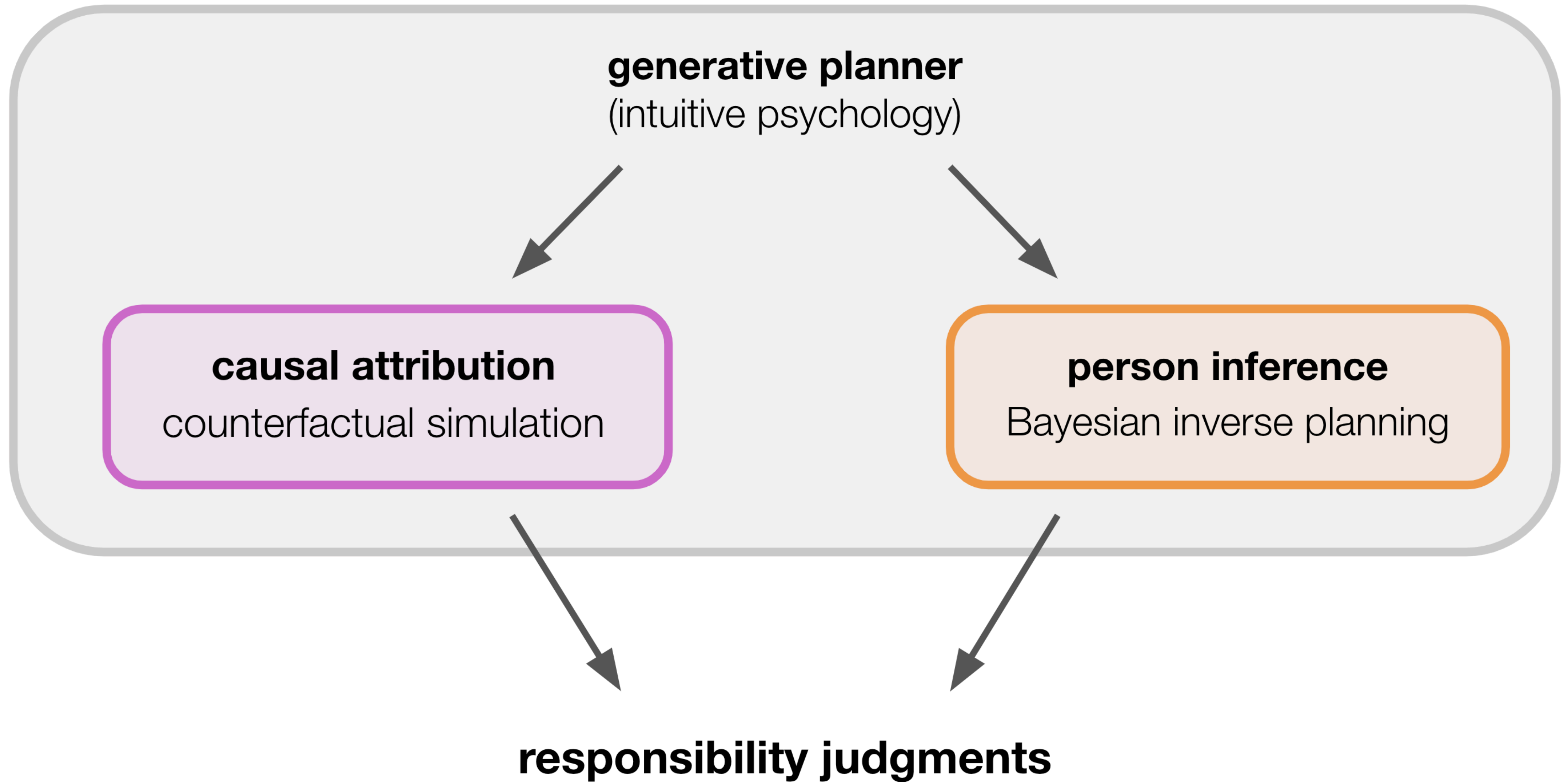
3

result:

SUCCESS

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



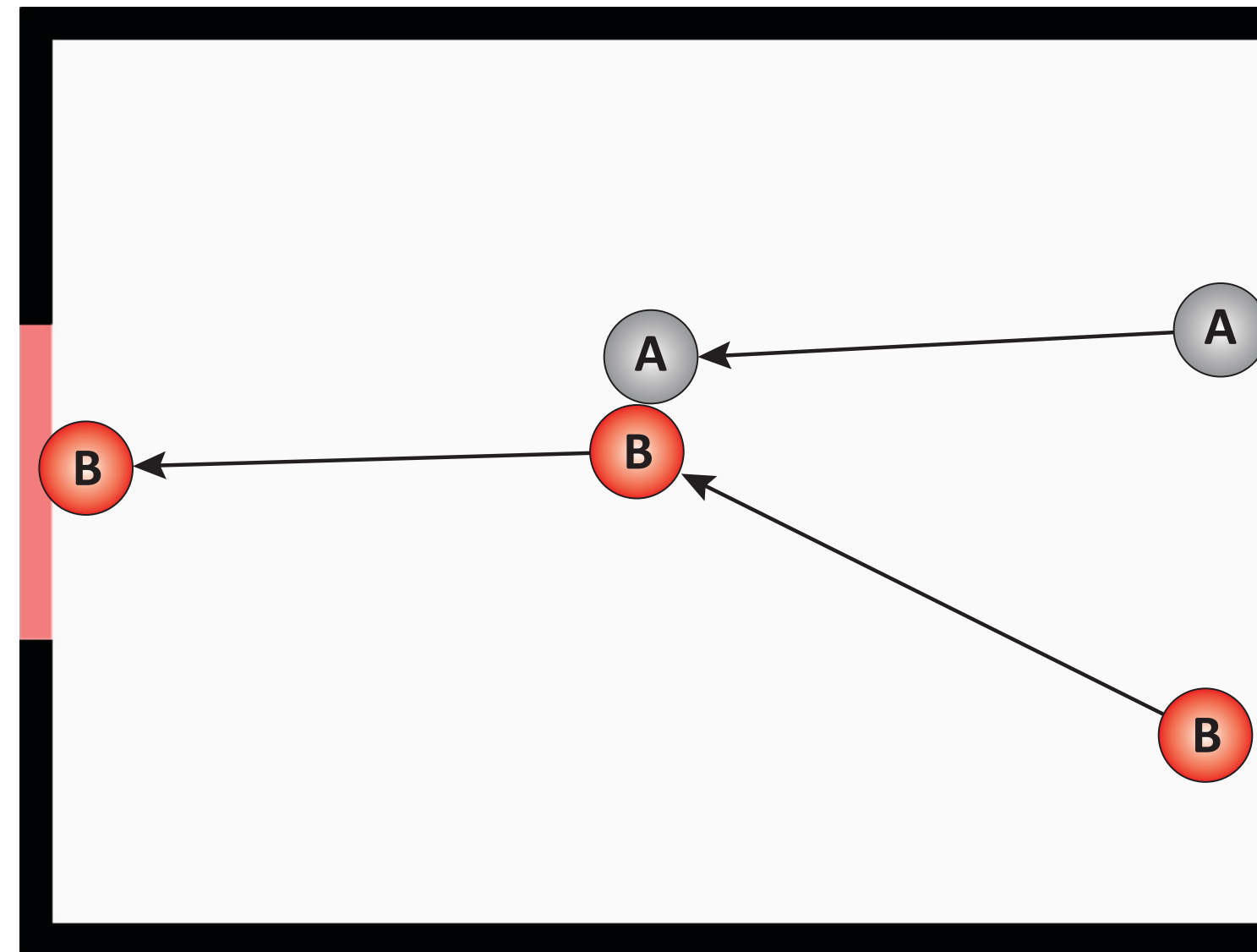


causal attribution

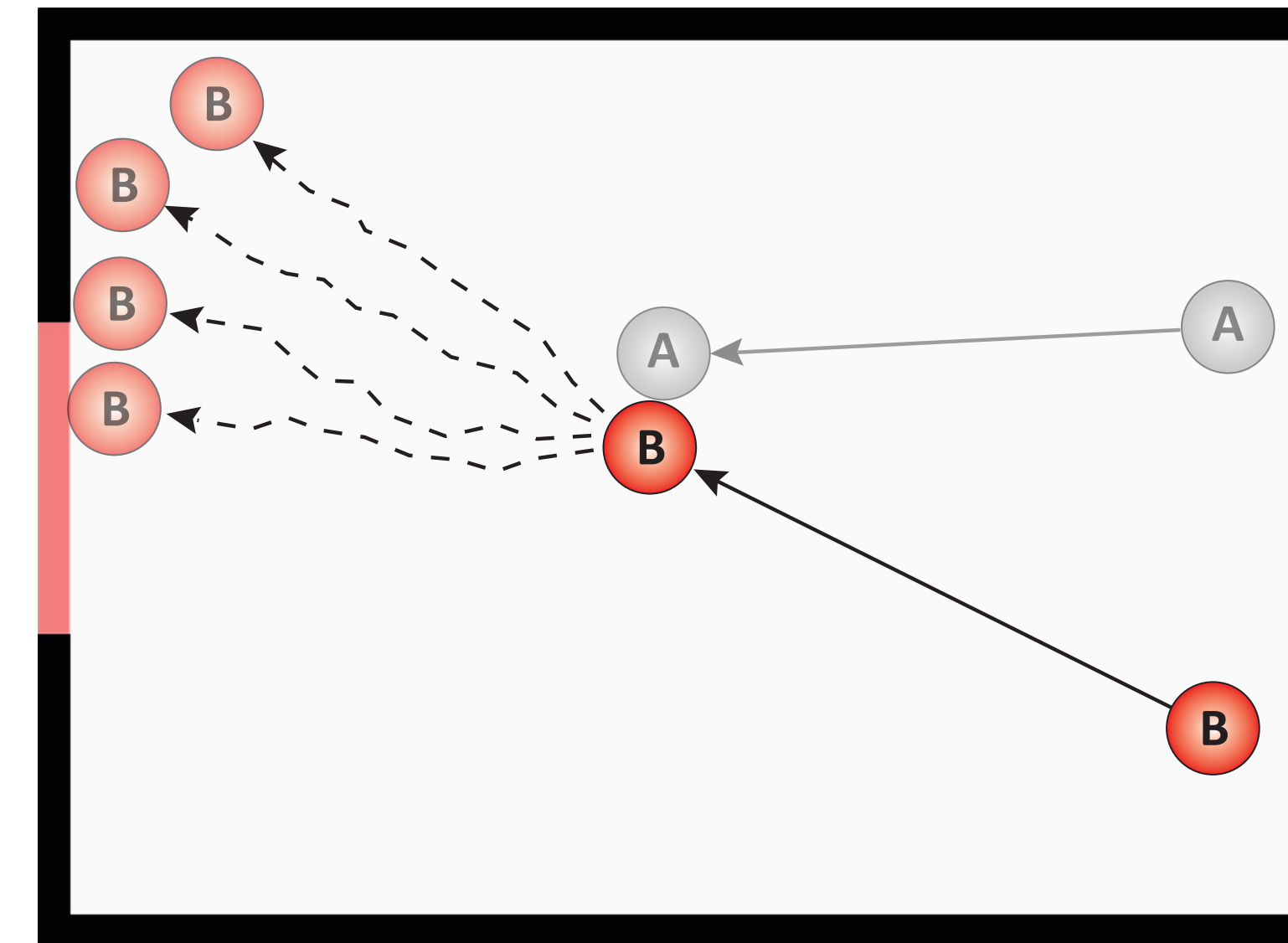
counterfactual simulation

intuitive
physics

actual situation



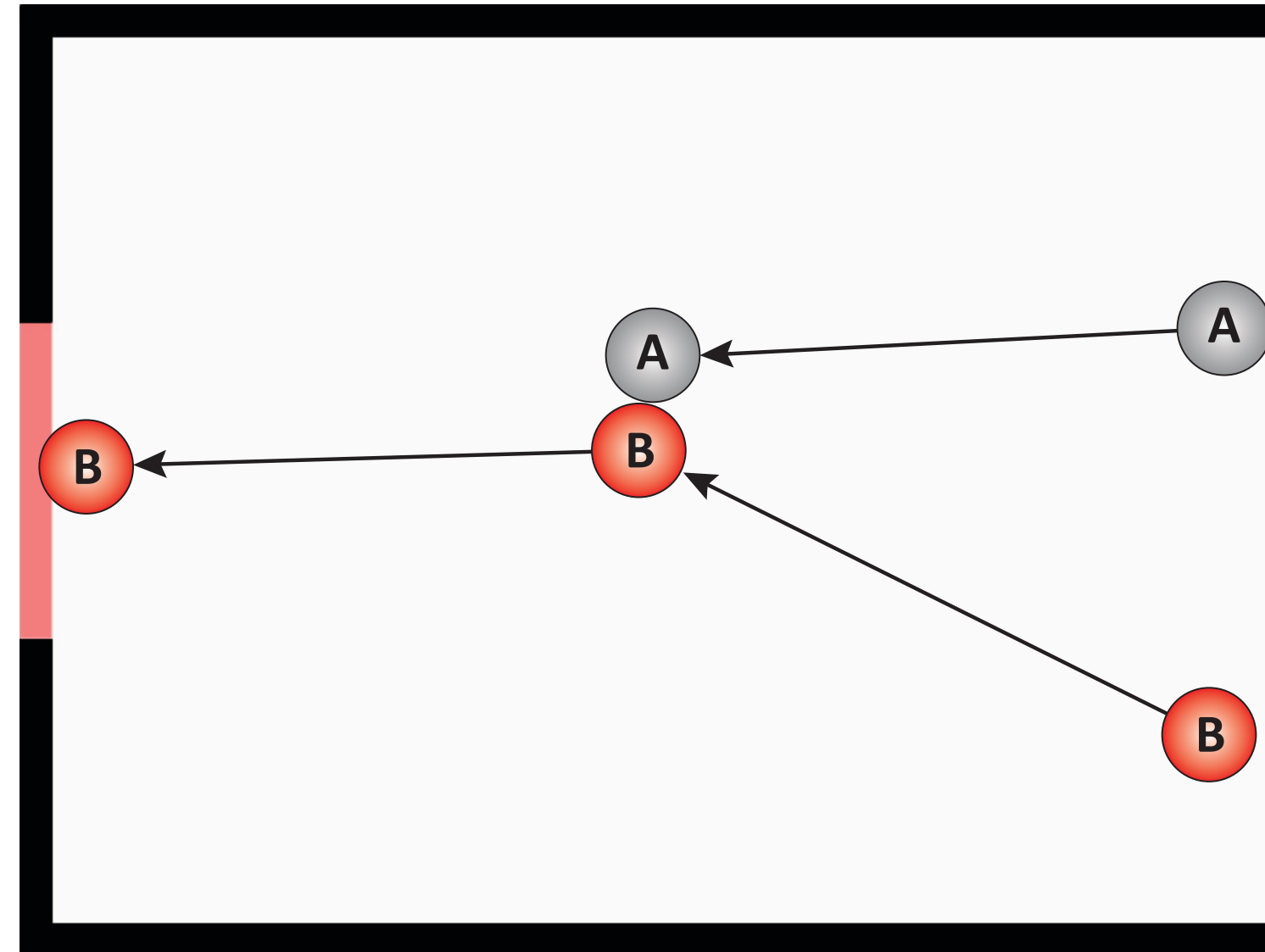
counterfactual simulations



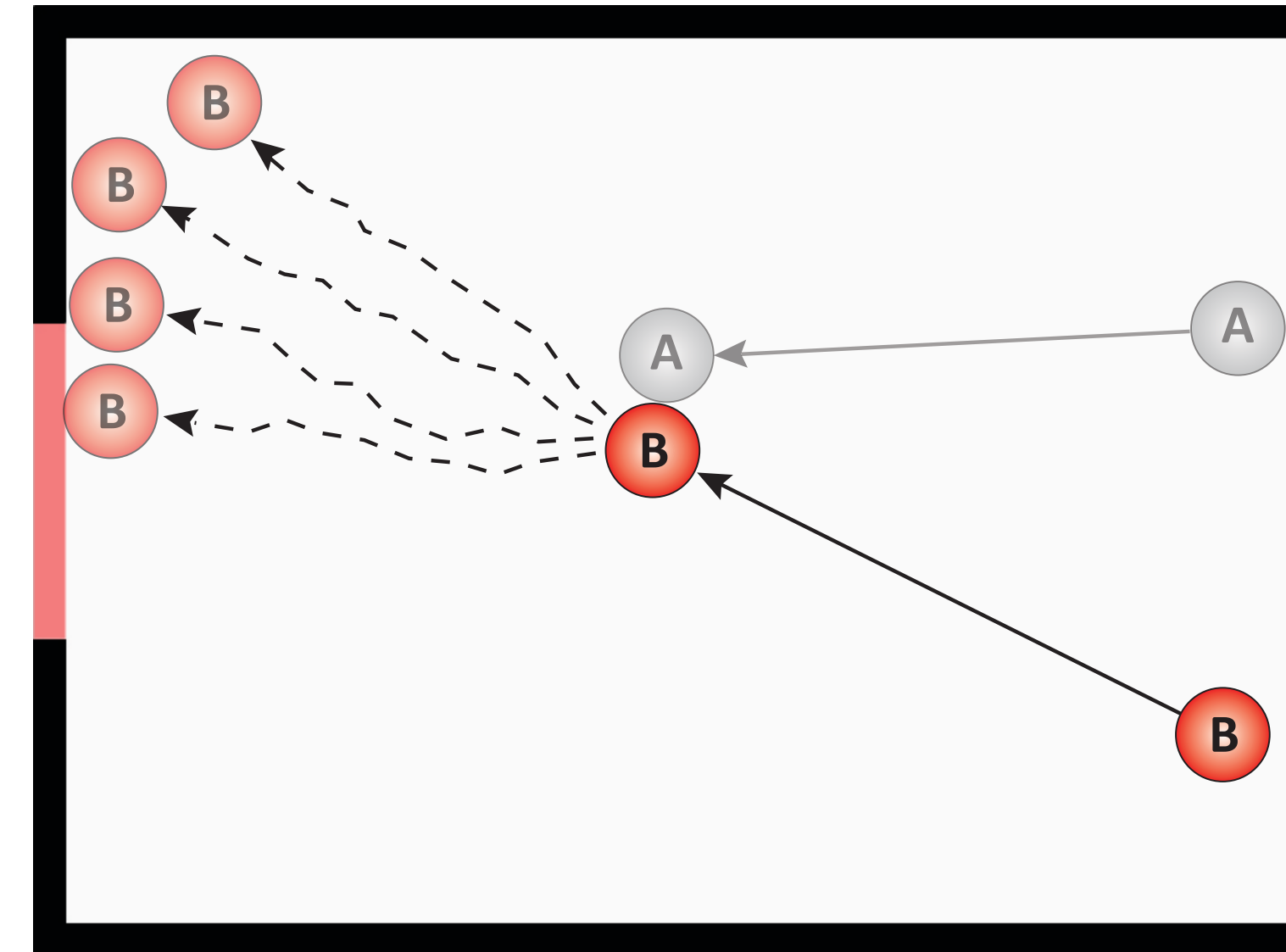
causal attribution
counterfactual simulation

intuitive
physics

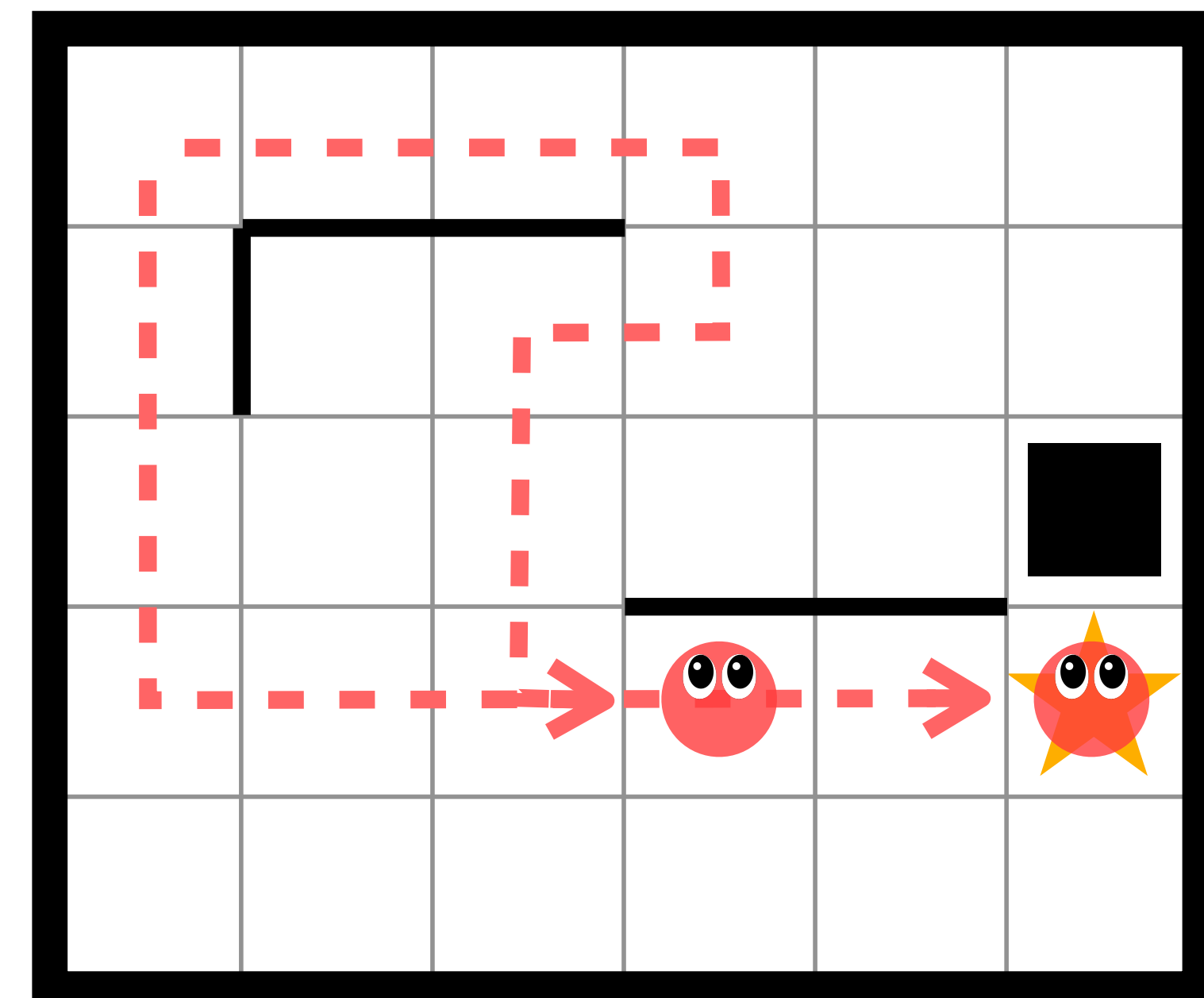
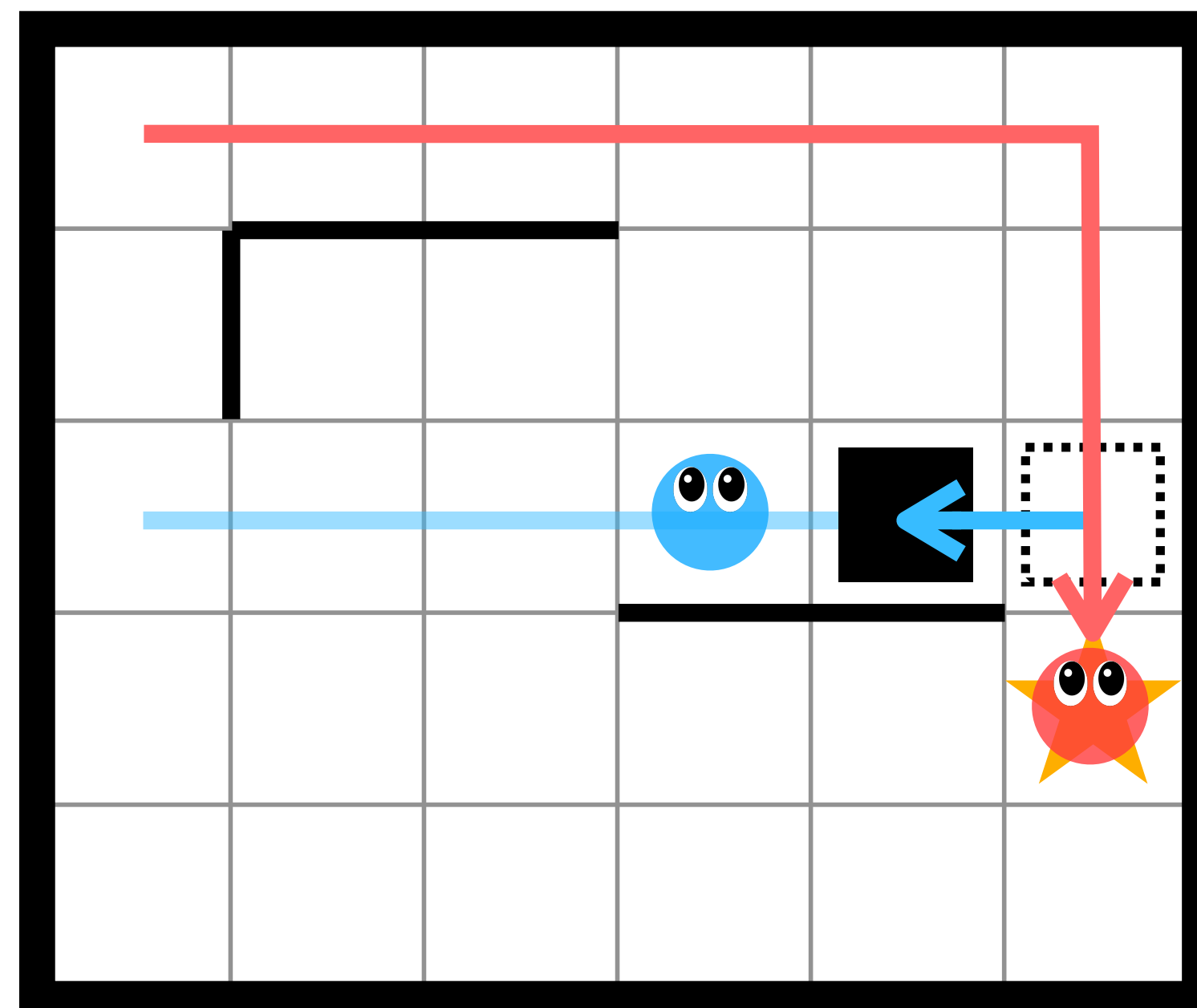
actual situation



counterfactual simulations



intuitive
psychology

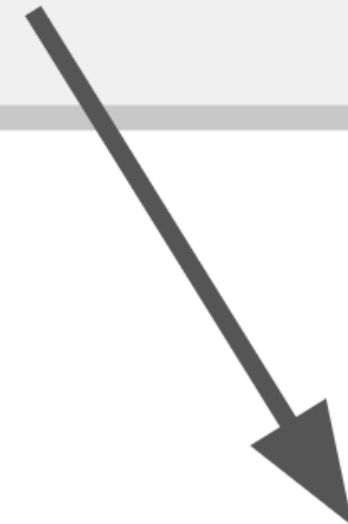


generative planner
(intuitive psychology)




causal attribution
counterfactual simulation


person inference
Bayesian inverse planning

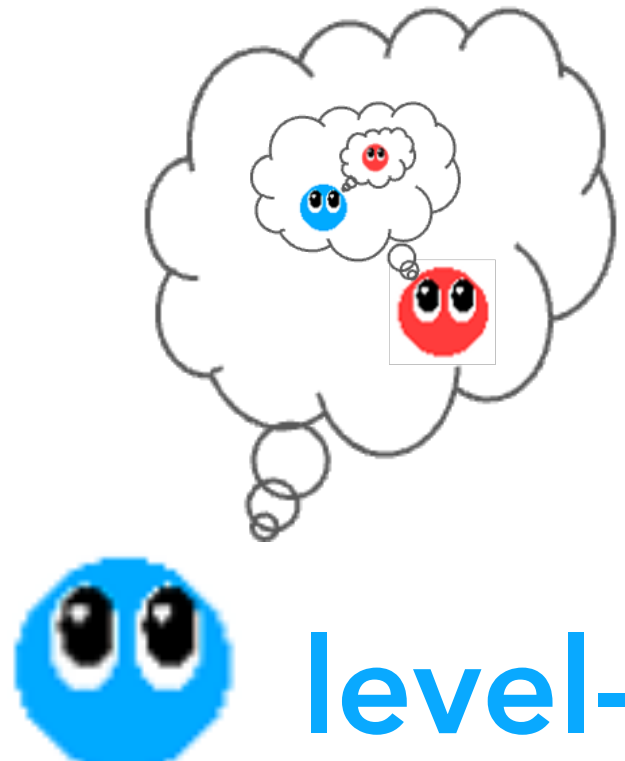


responsibility judgments

 **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

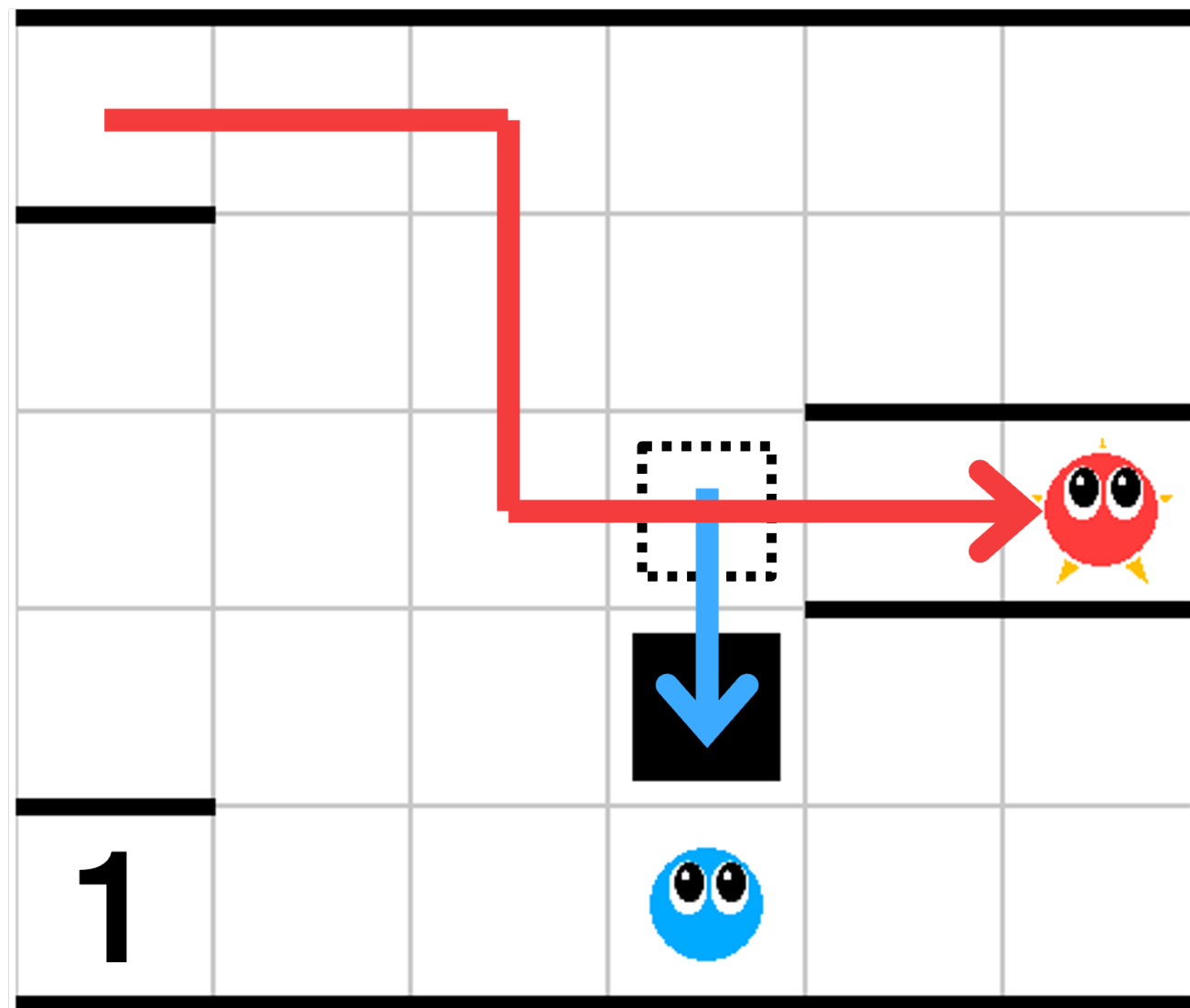
 **level-3 blue** plans to help or deceive a **level-2 red**

causal attribution

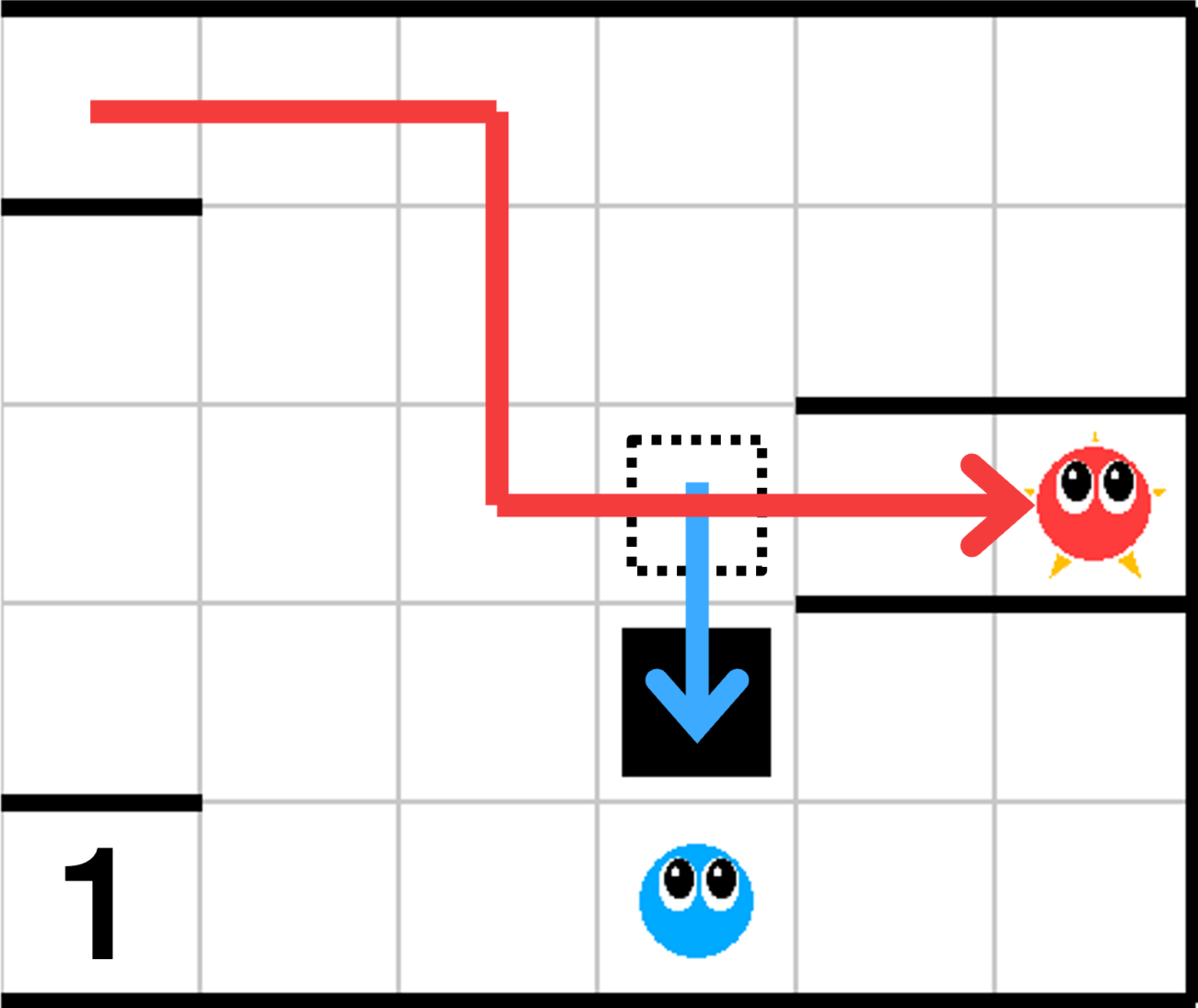
counterfactual simulation

person inference

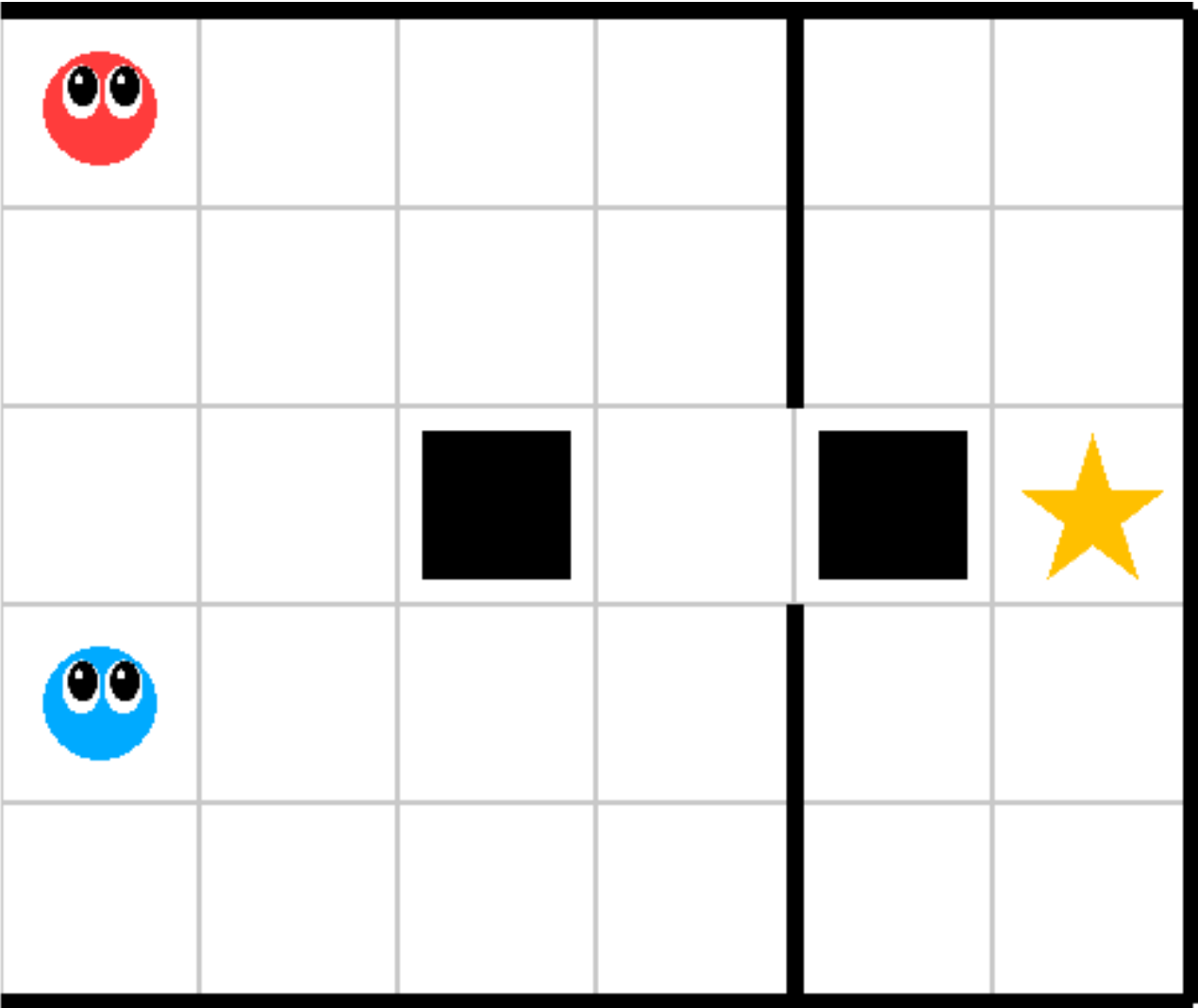
Bayesian inverse planning



causal attribution
counterfactual simulation



person inference
Bayesian inverse planning

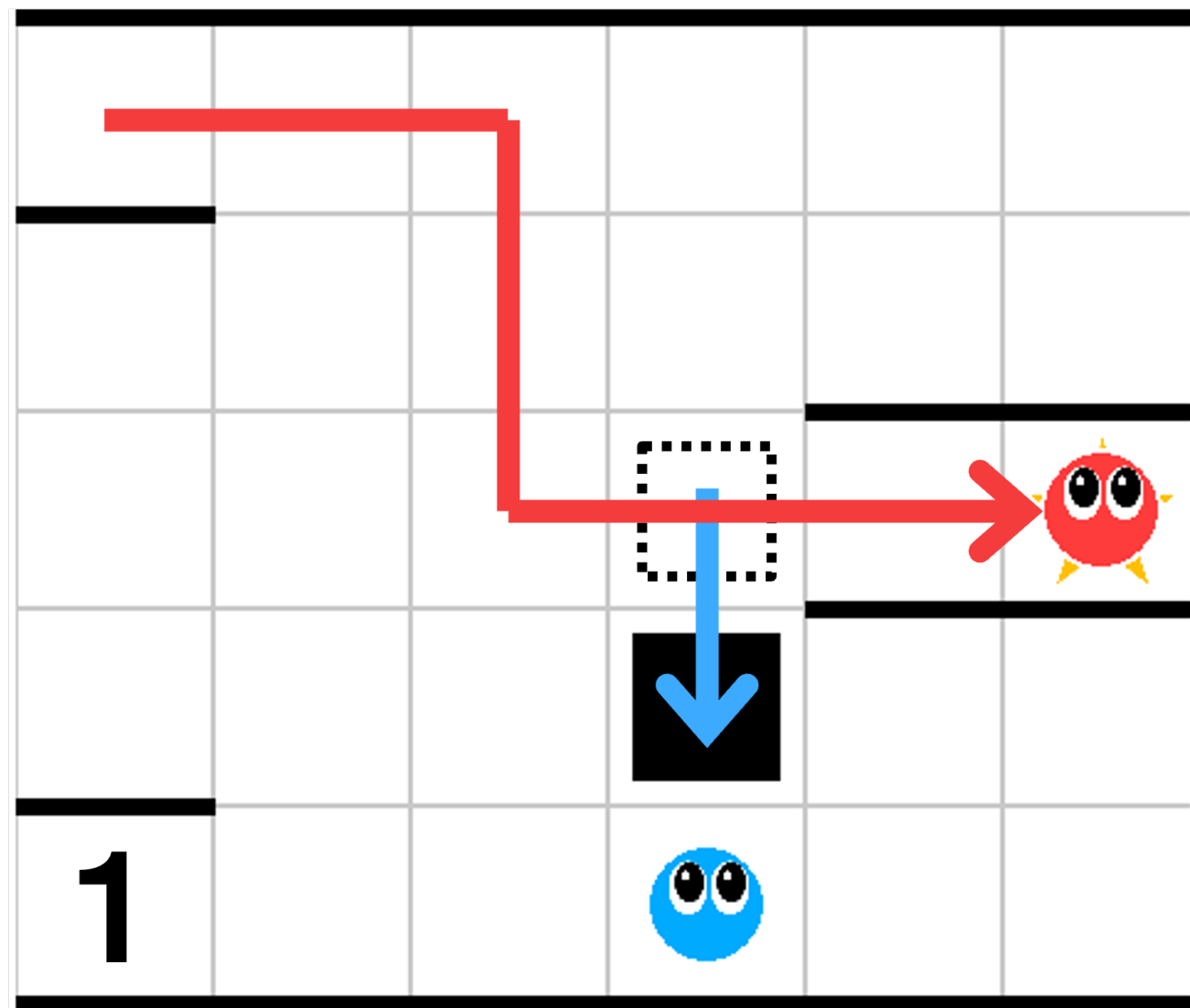


time left:
10

result:

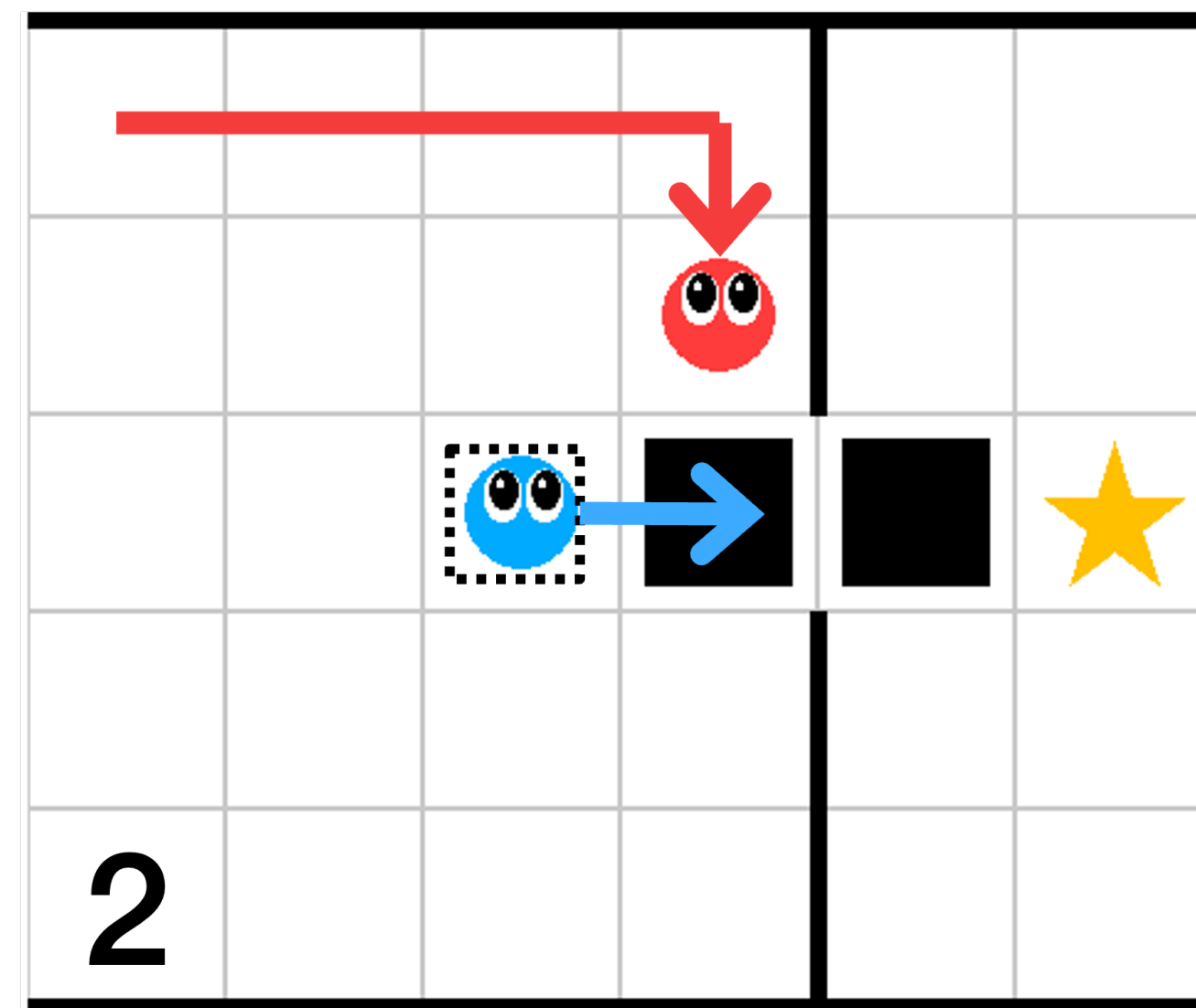
causal attribution

counterfactual simulation



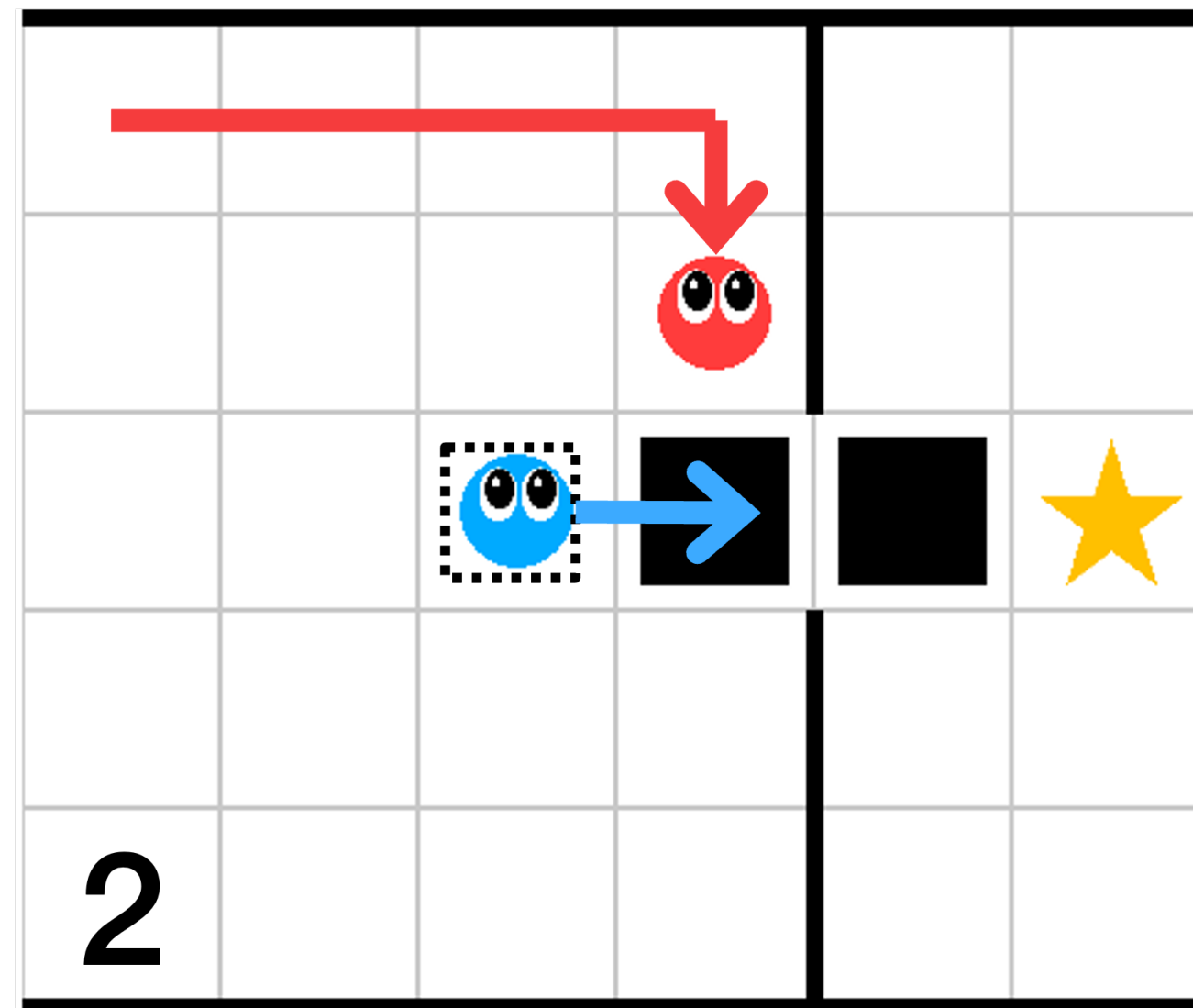
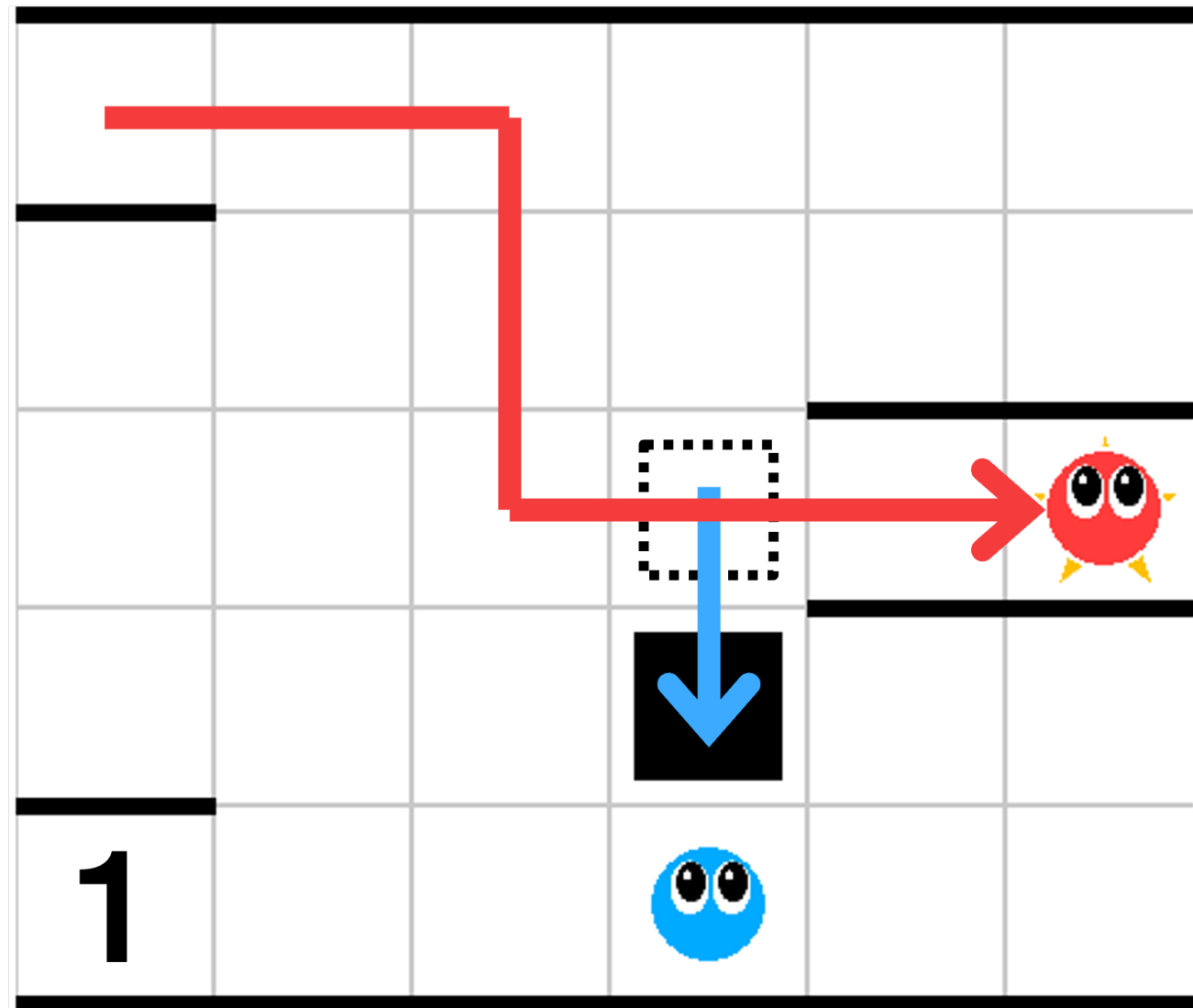
person inference

Bayesian inverse planning



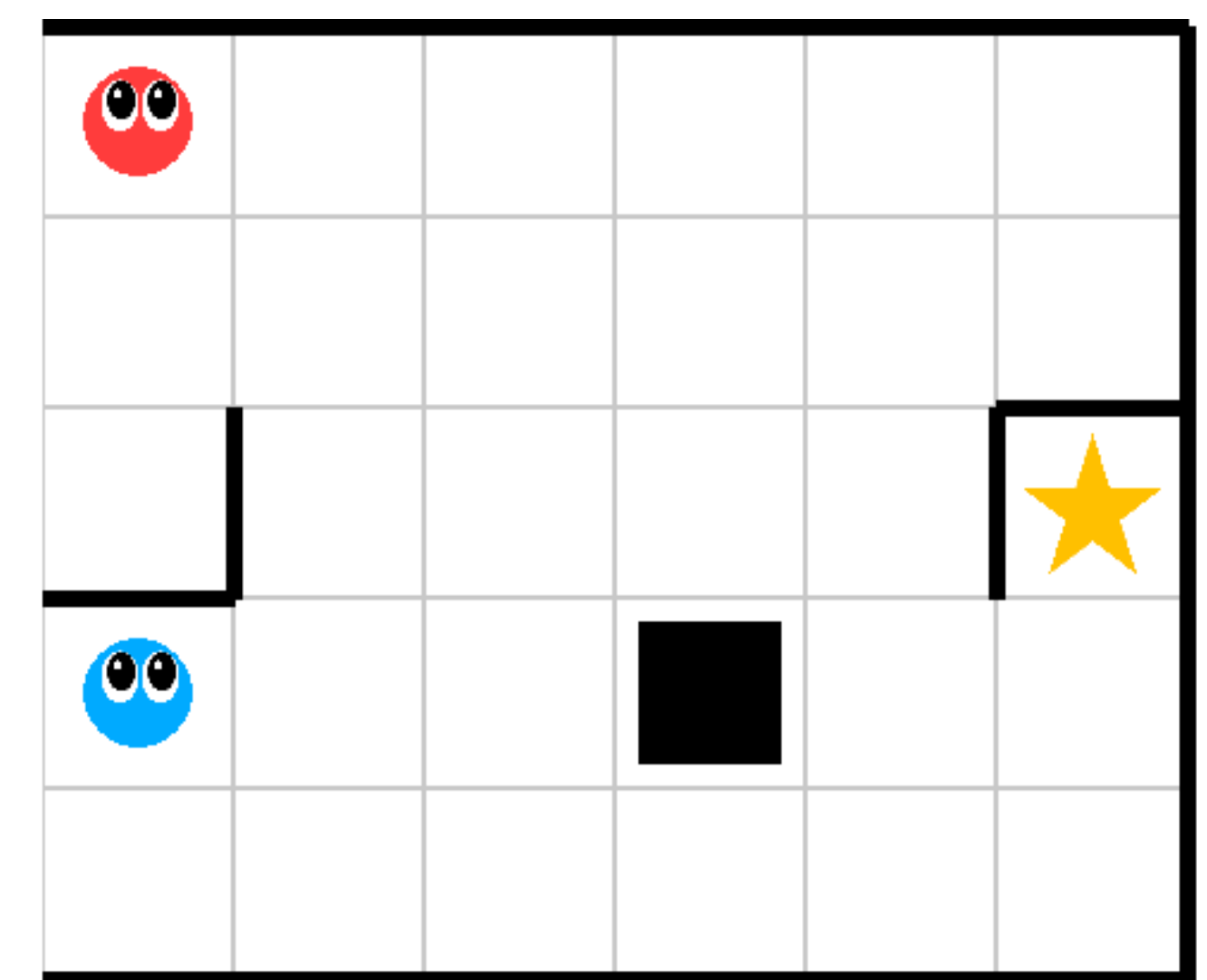
causal attribution

counterfactual simulation



person inference

Bayesian inverse planning

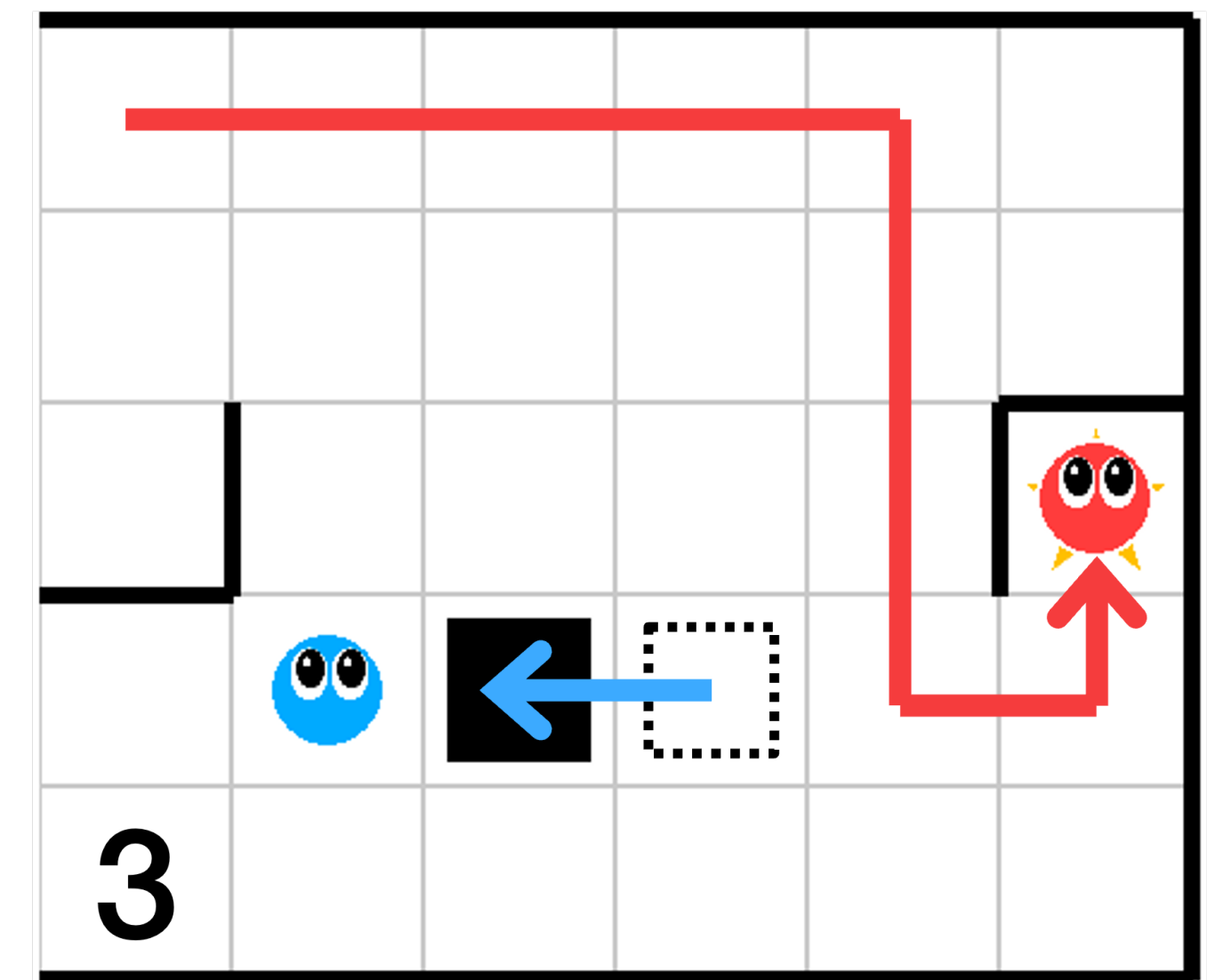
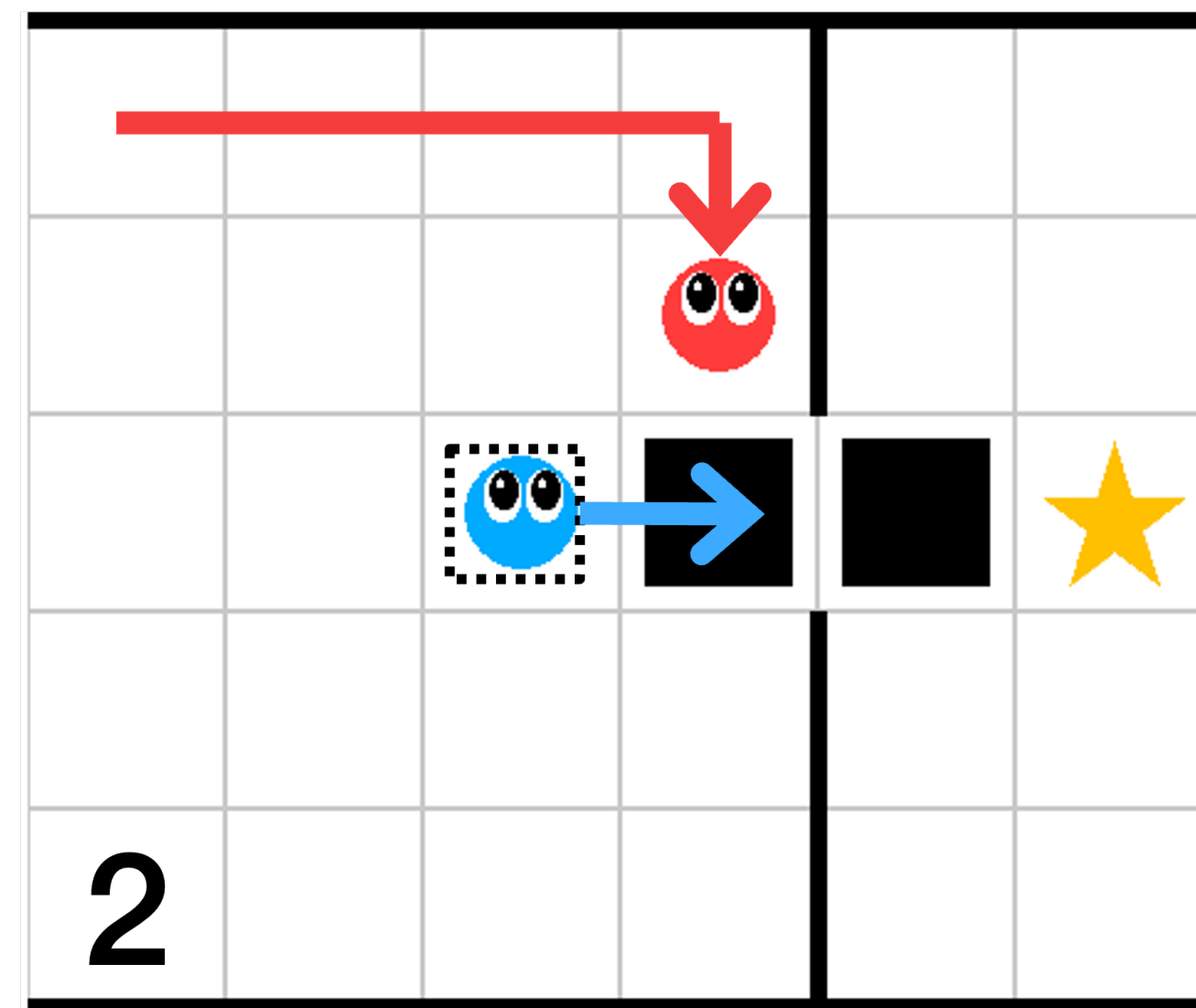
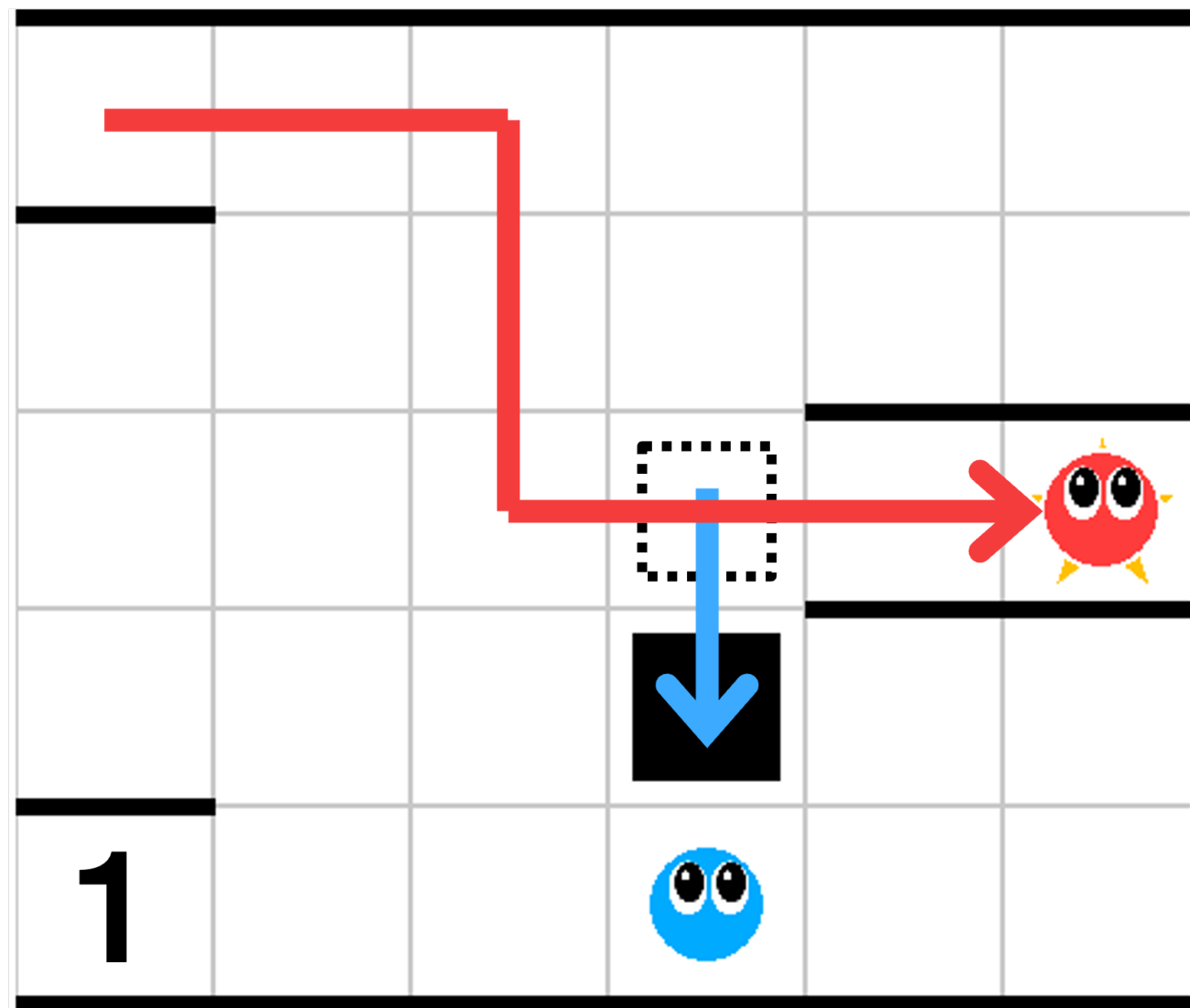


time left:
10

result:

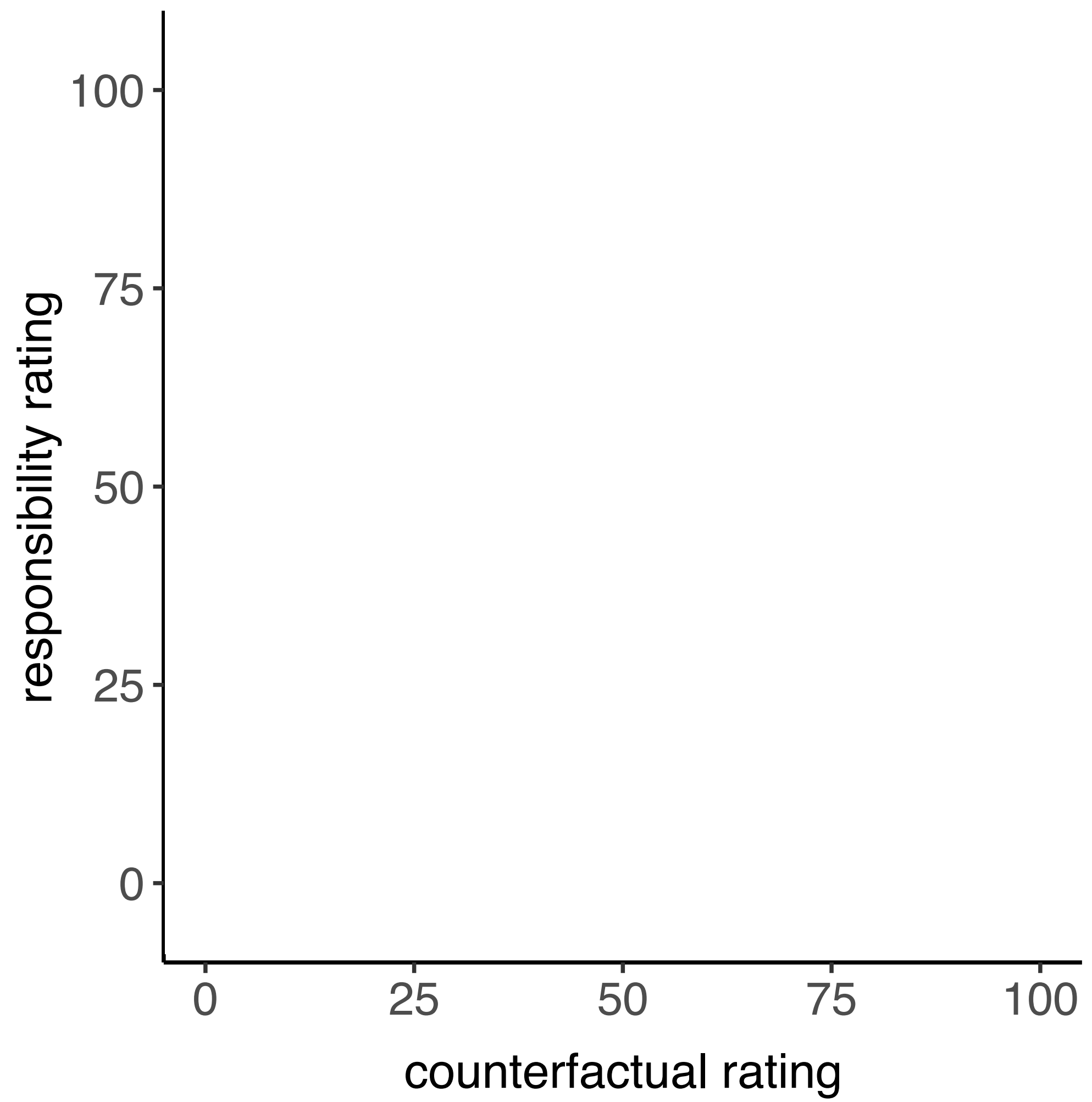
causal attribution

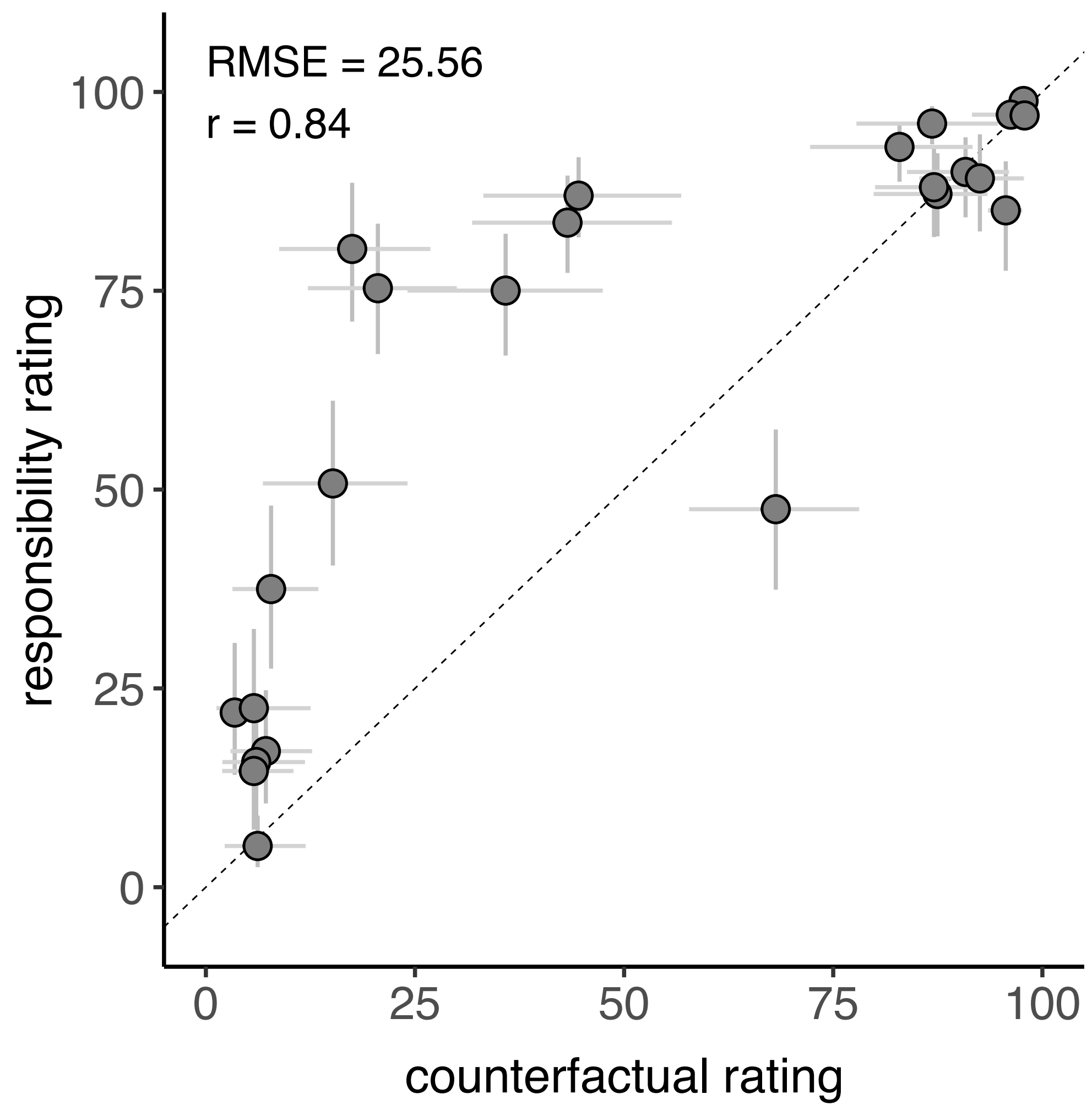
counterfactual simulation

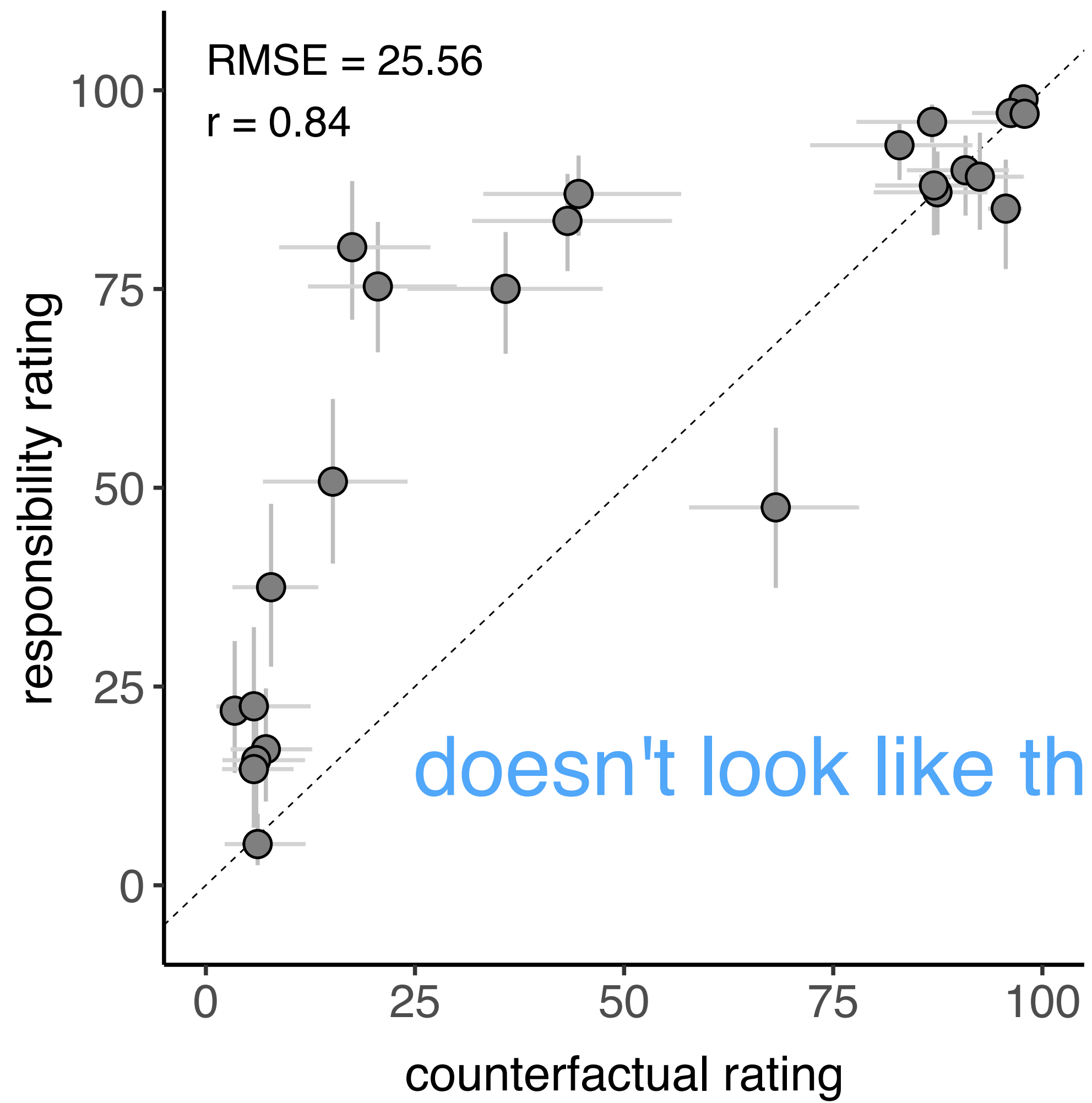


person inference

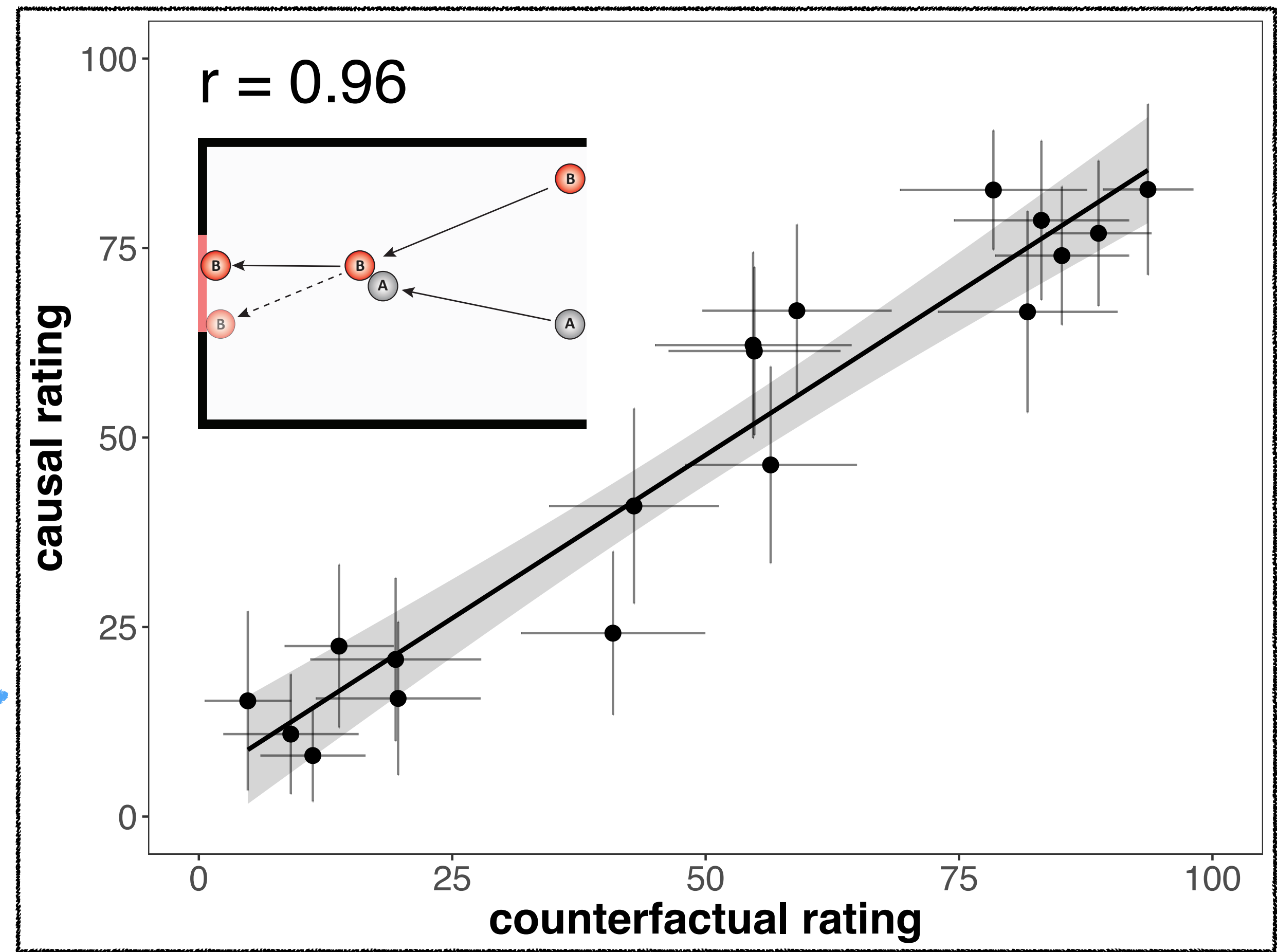
Bayesian inverse planning

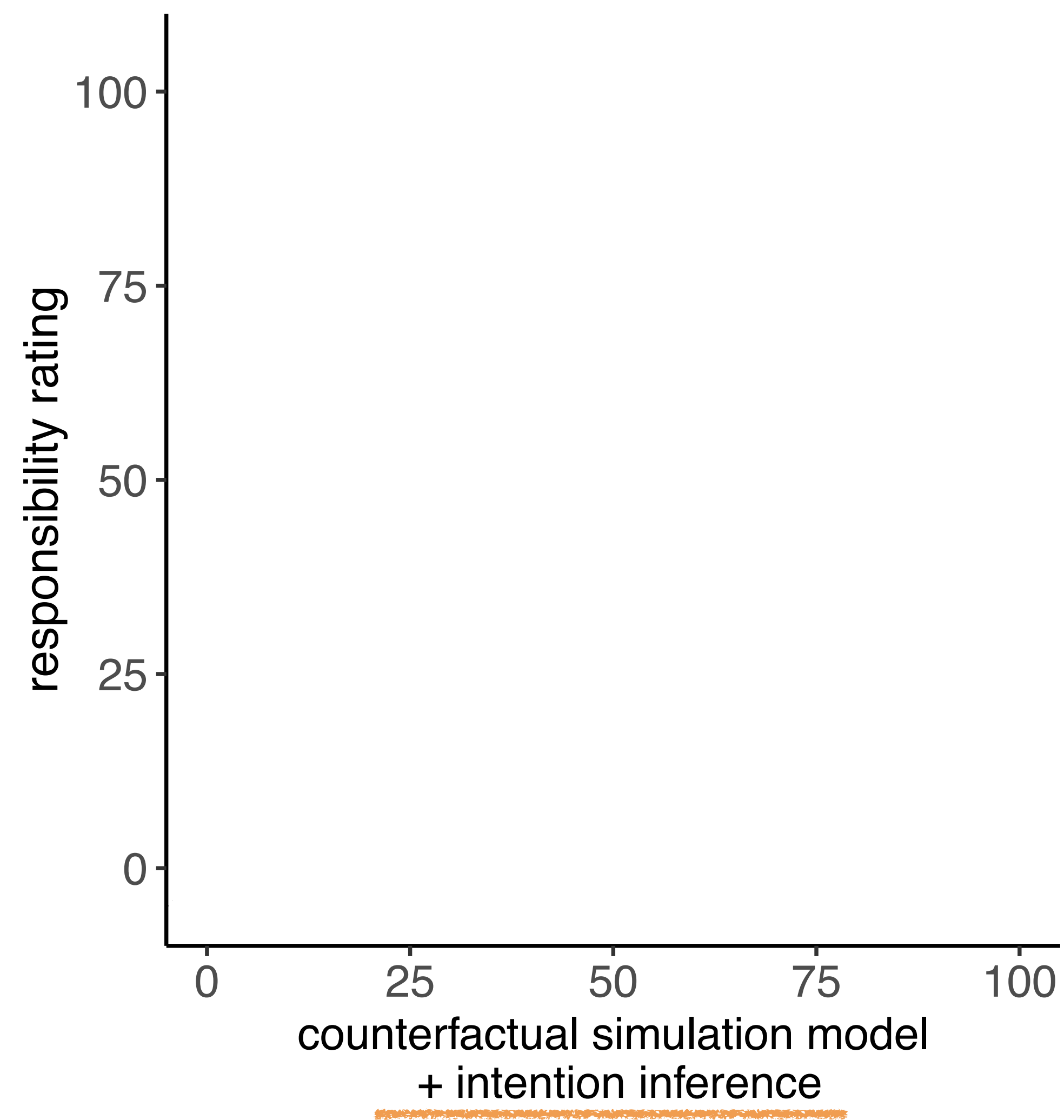
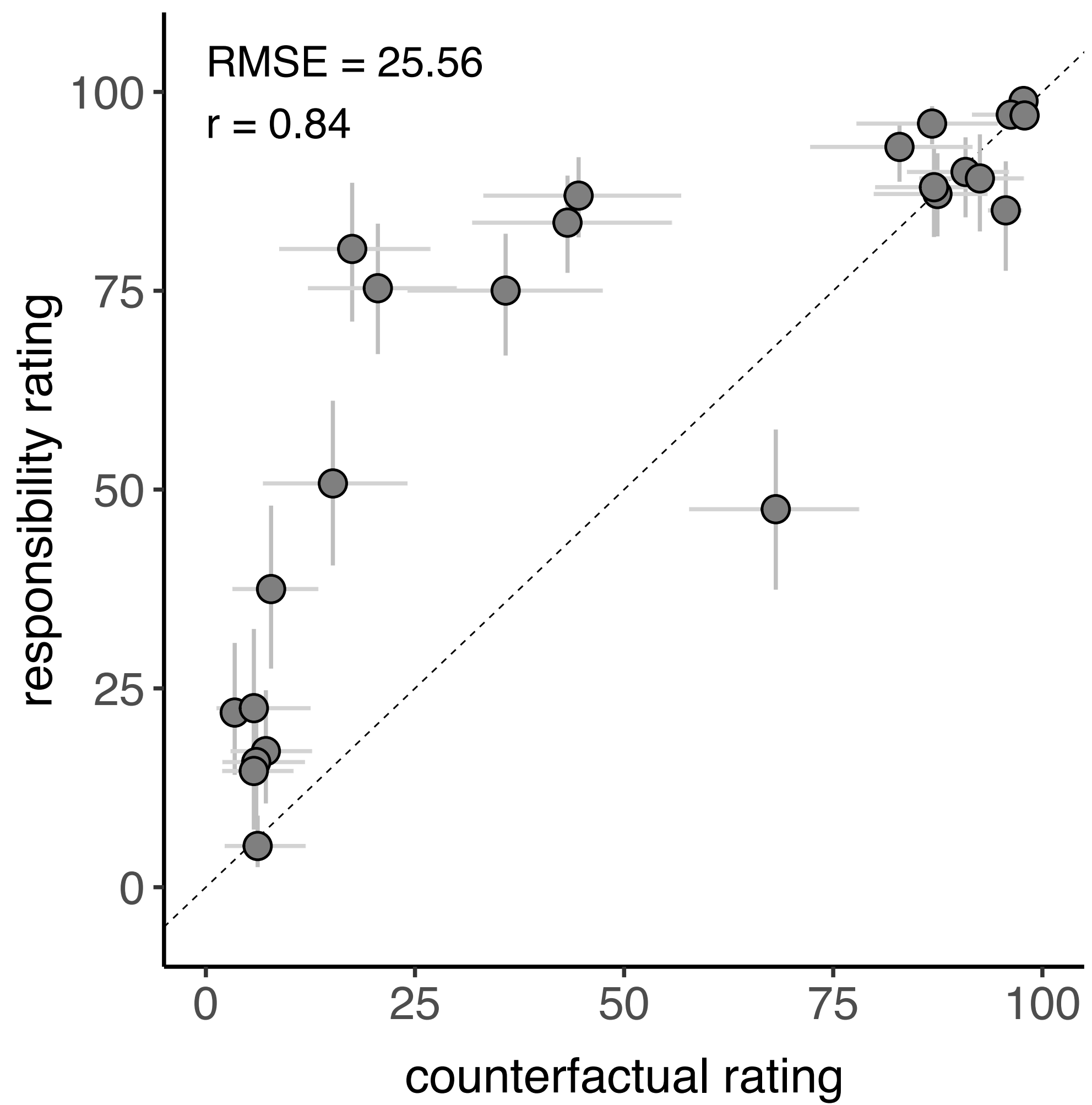


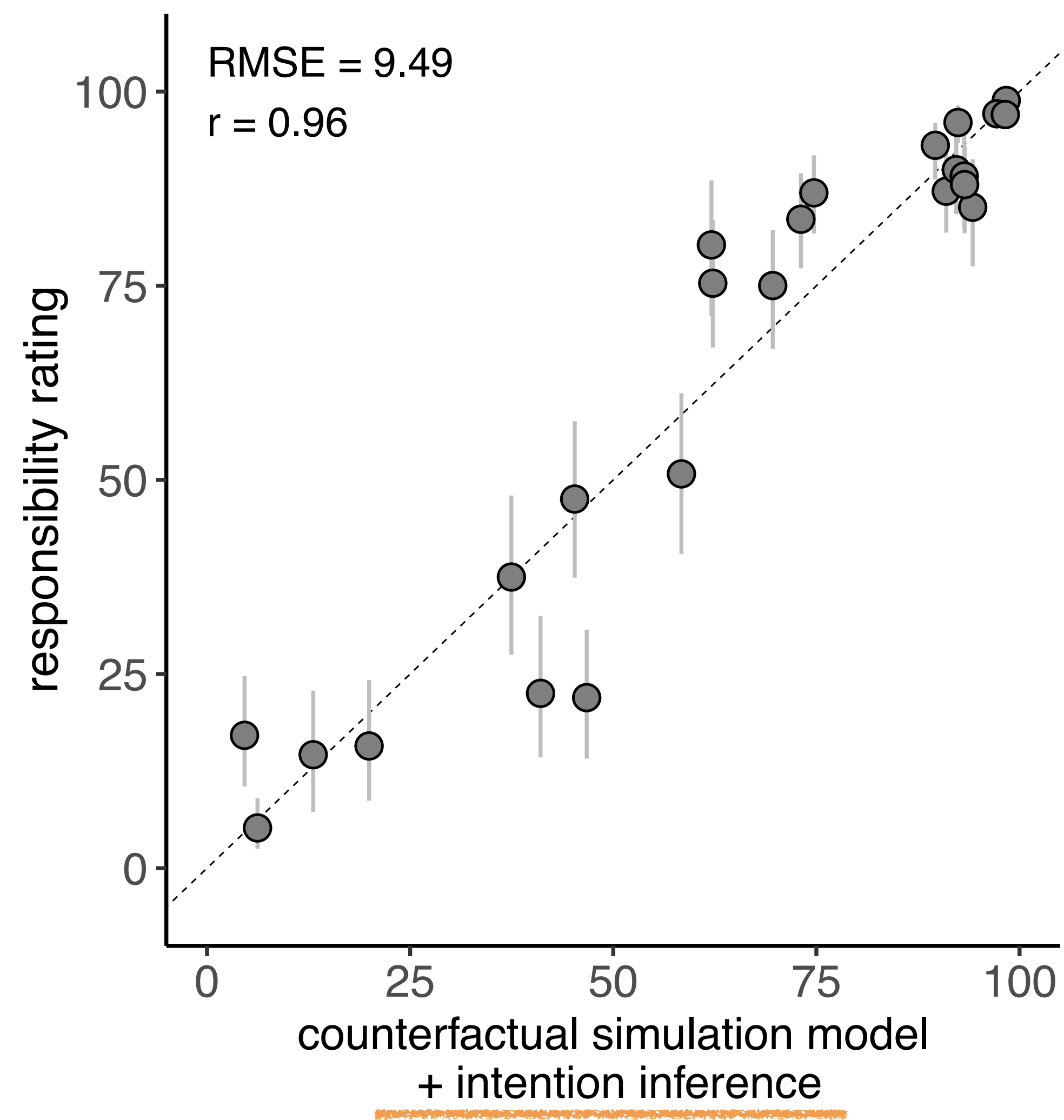
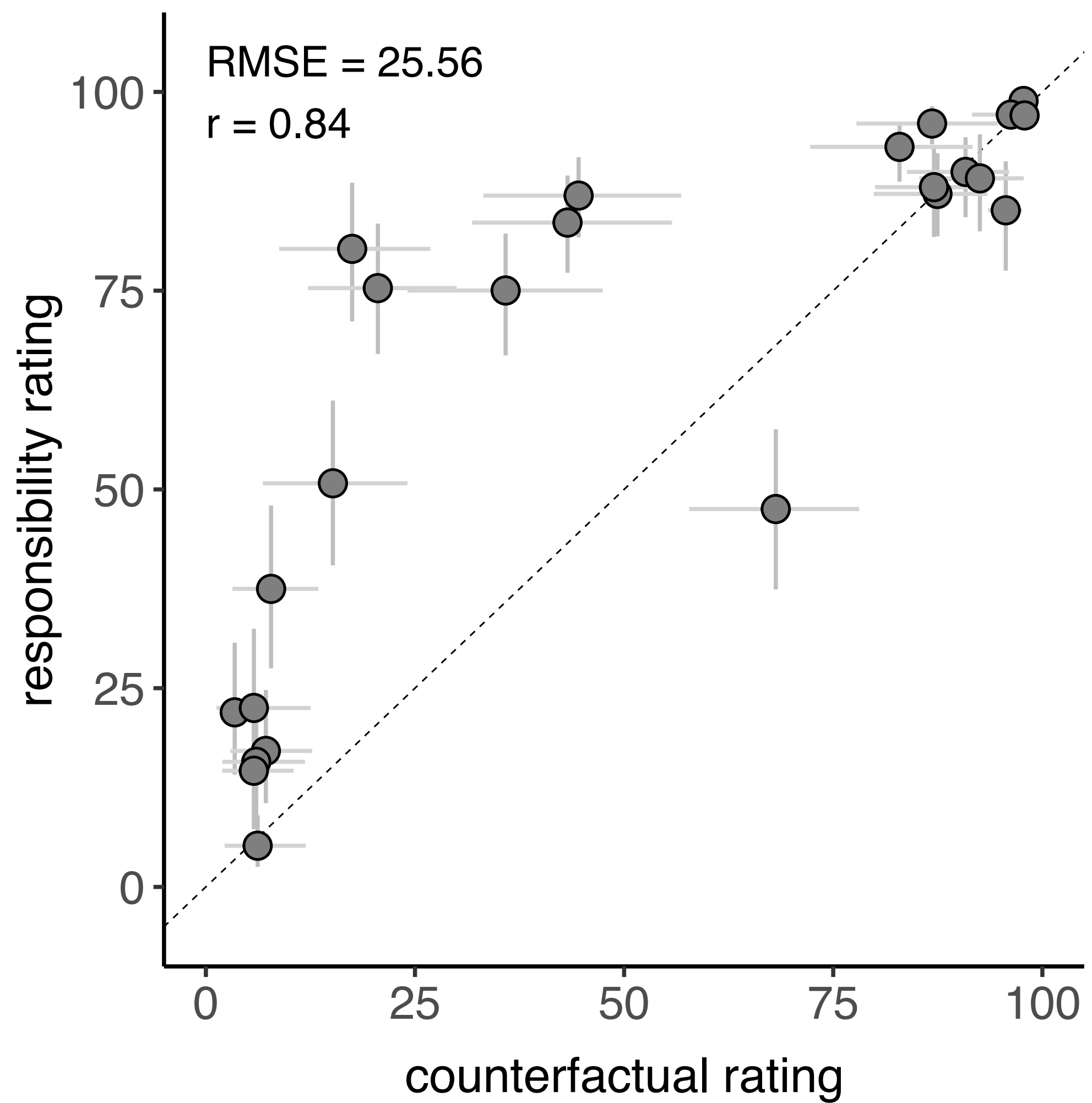


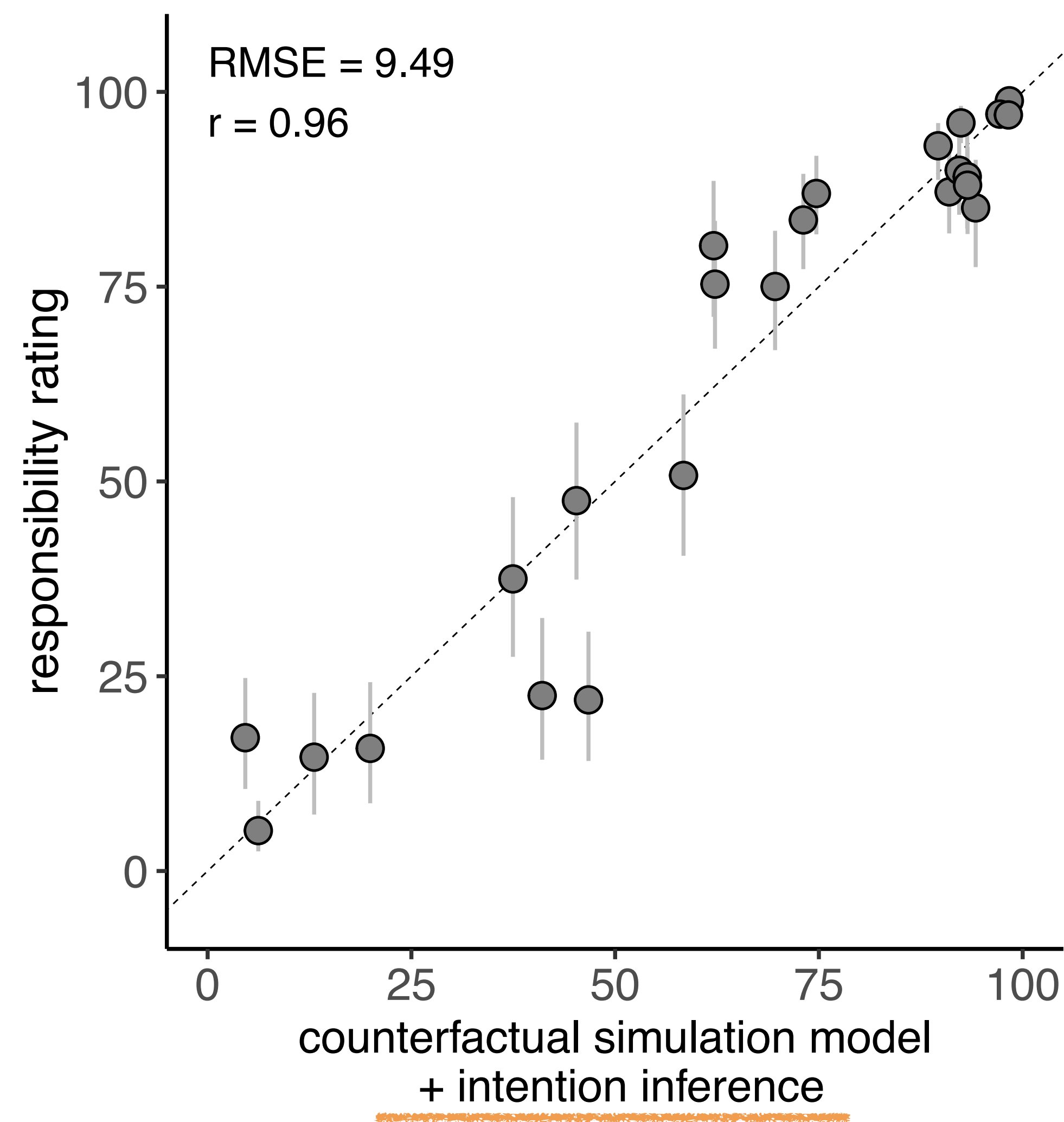
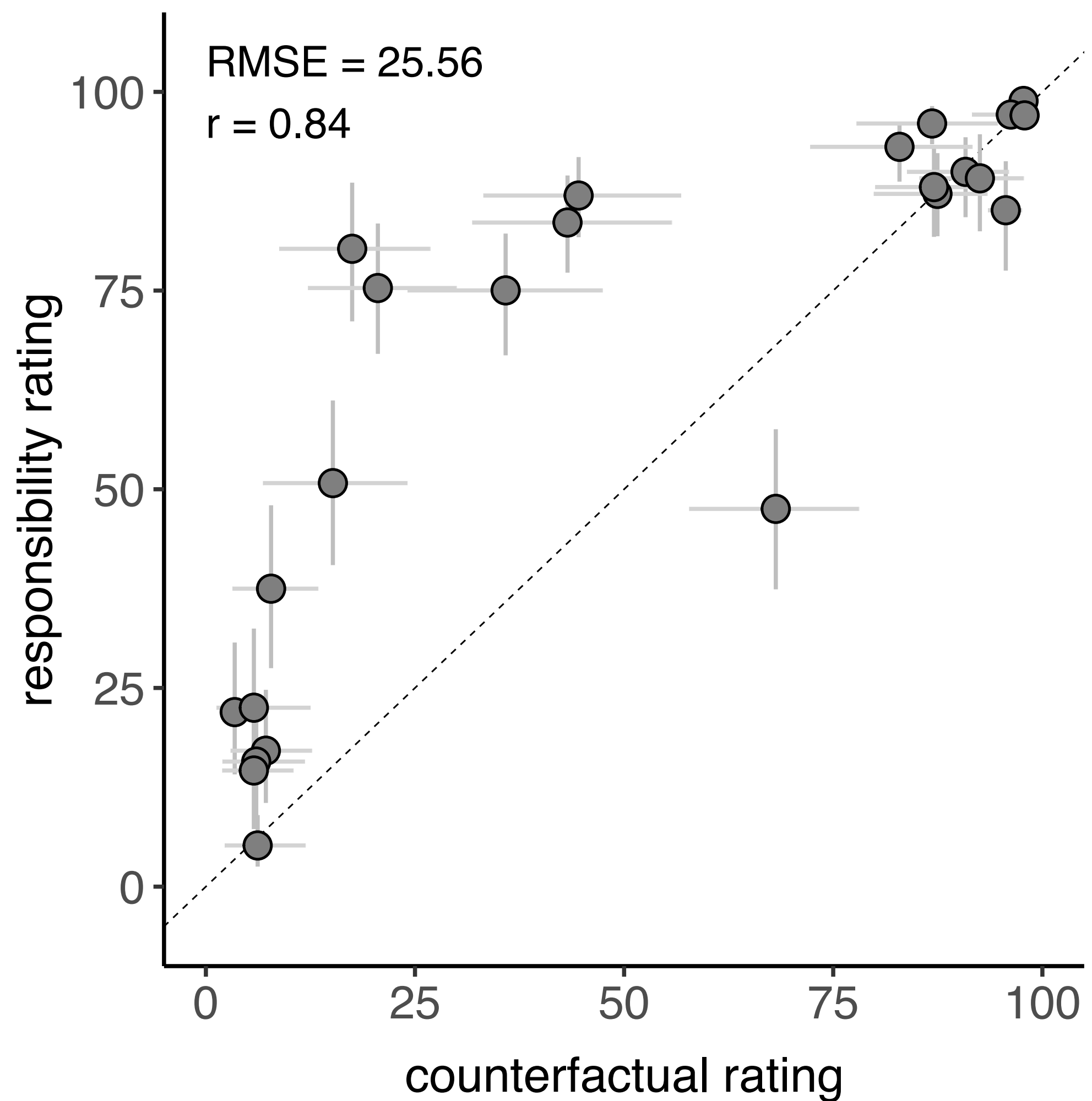


doesn't look like this →





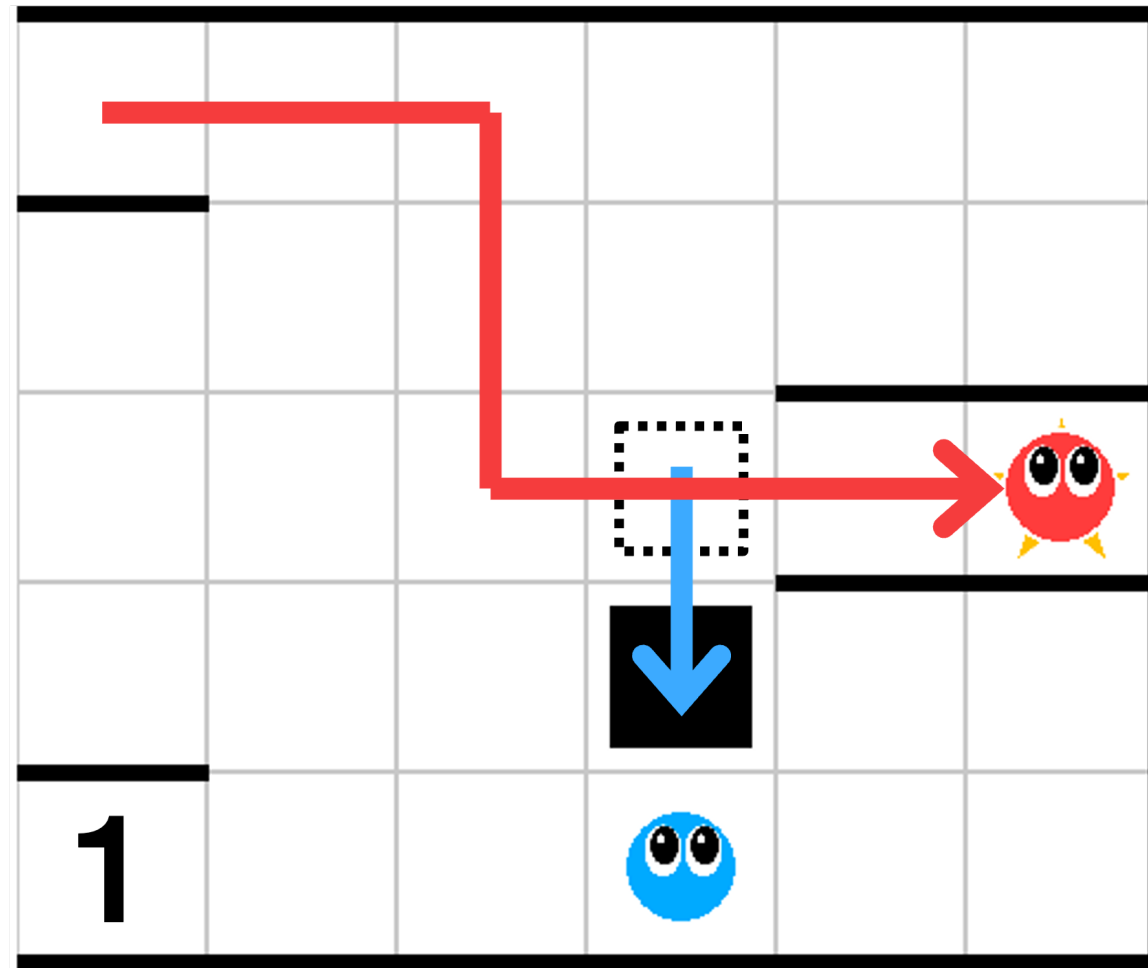




A model that combines
counterfactual simulation + intention inference
accurately captures responsibility judgments

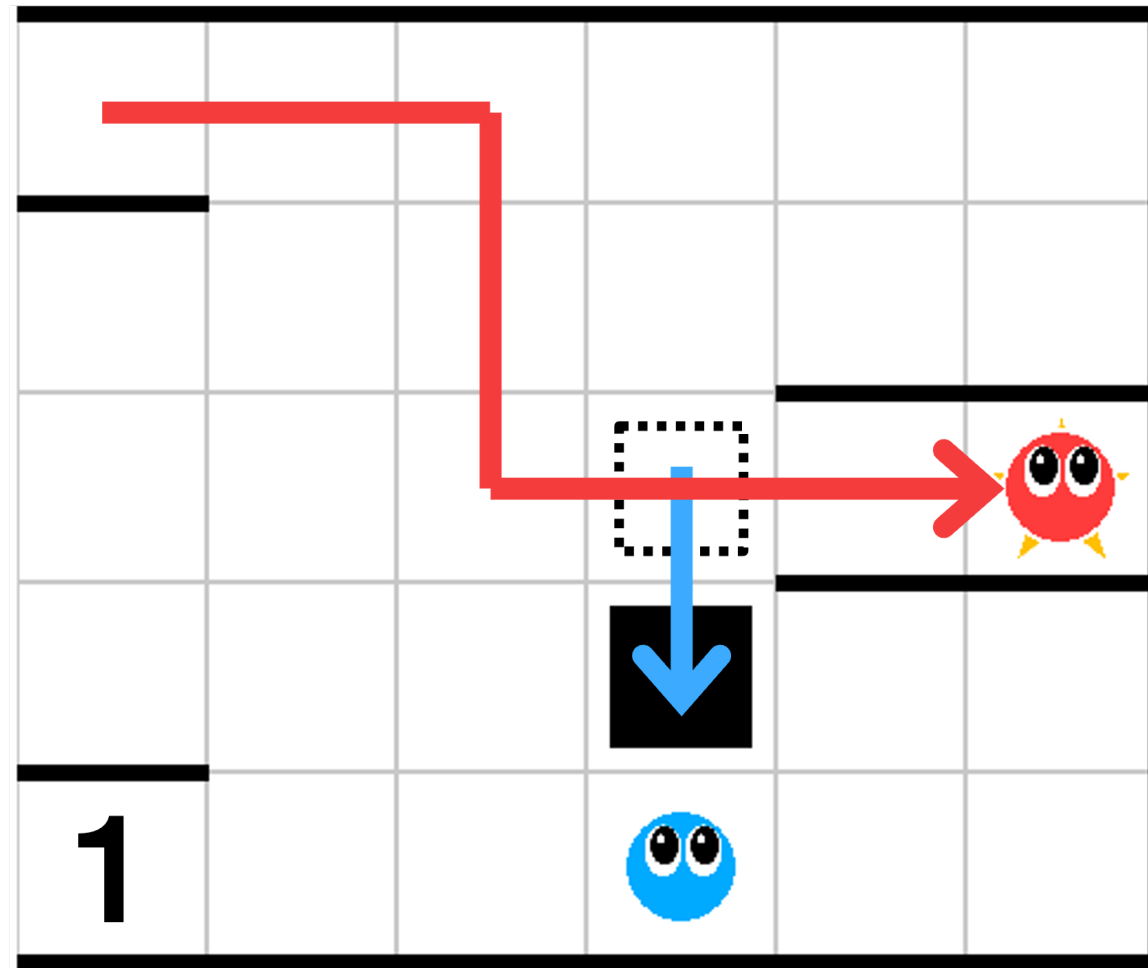
Counterfactual simulation & intuitive psychology

Counterfactual simulation & intuitive psychology

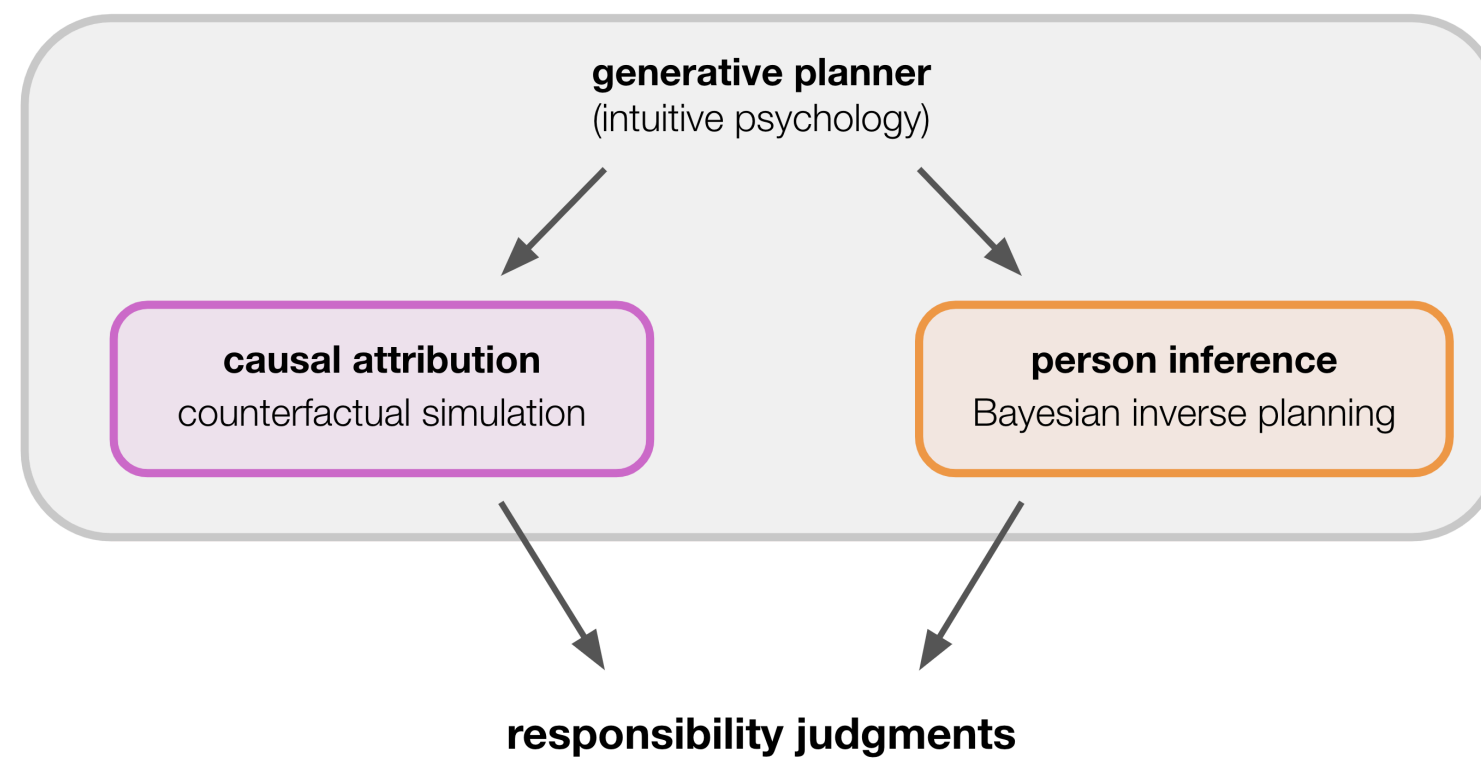


Judging whether someone **helped or hindered** requires counterfactual simulation

Counterfactual simulation & intuitive psychology



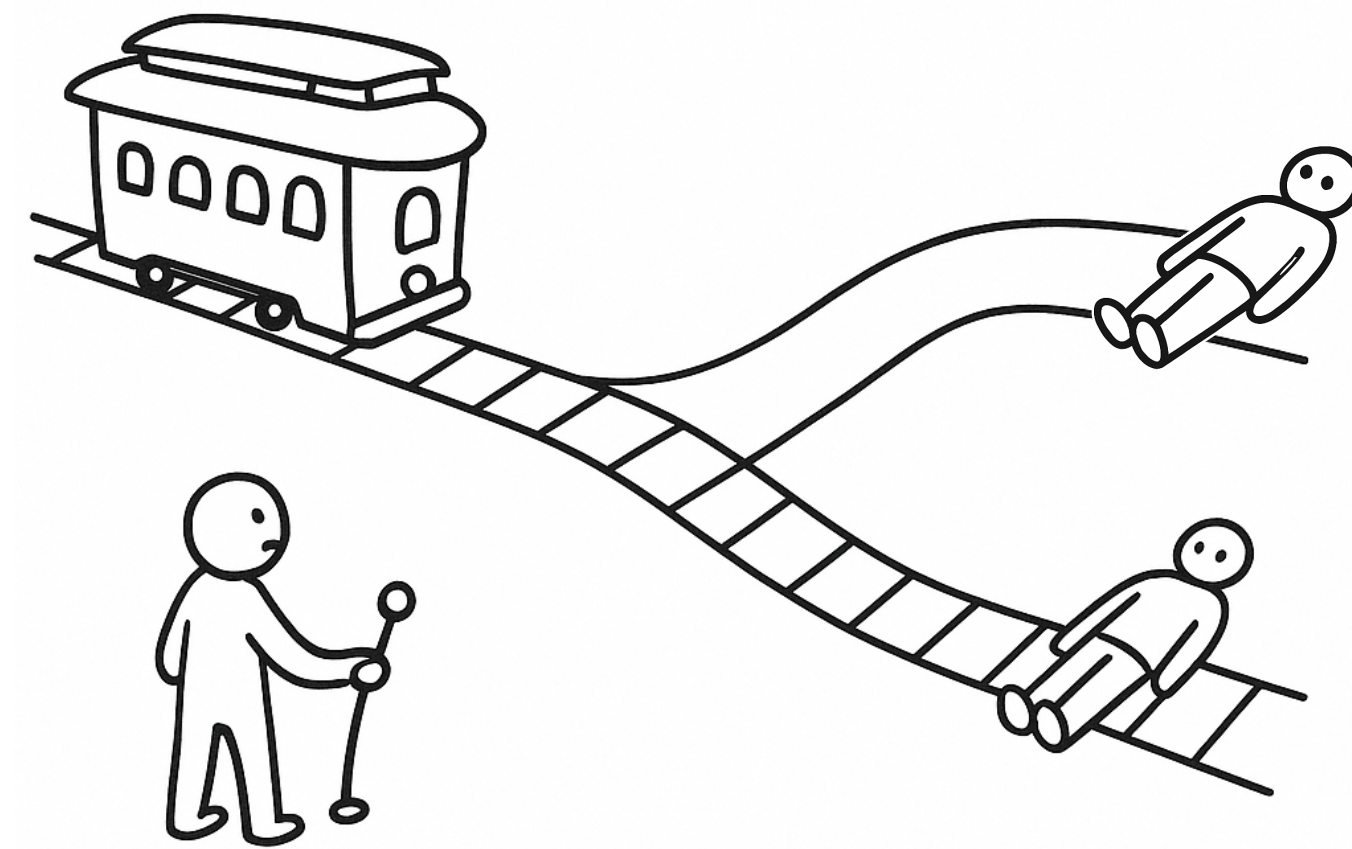
Judging whether someone **helped or hindered** requires counterfactual simulation



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

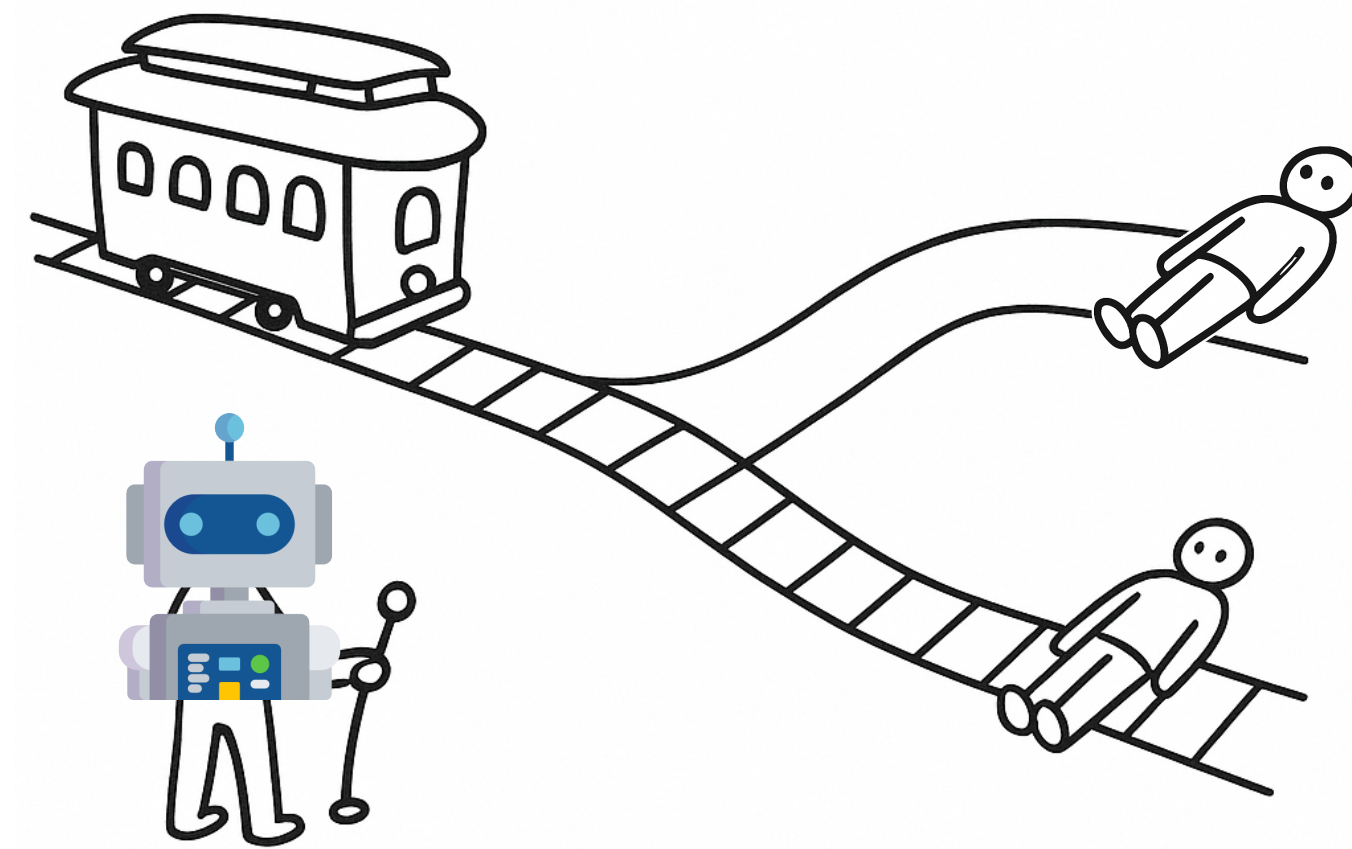
Are counterfactuals relevant for AI?

Are counterfactuals relevant for AI?

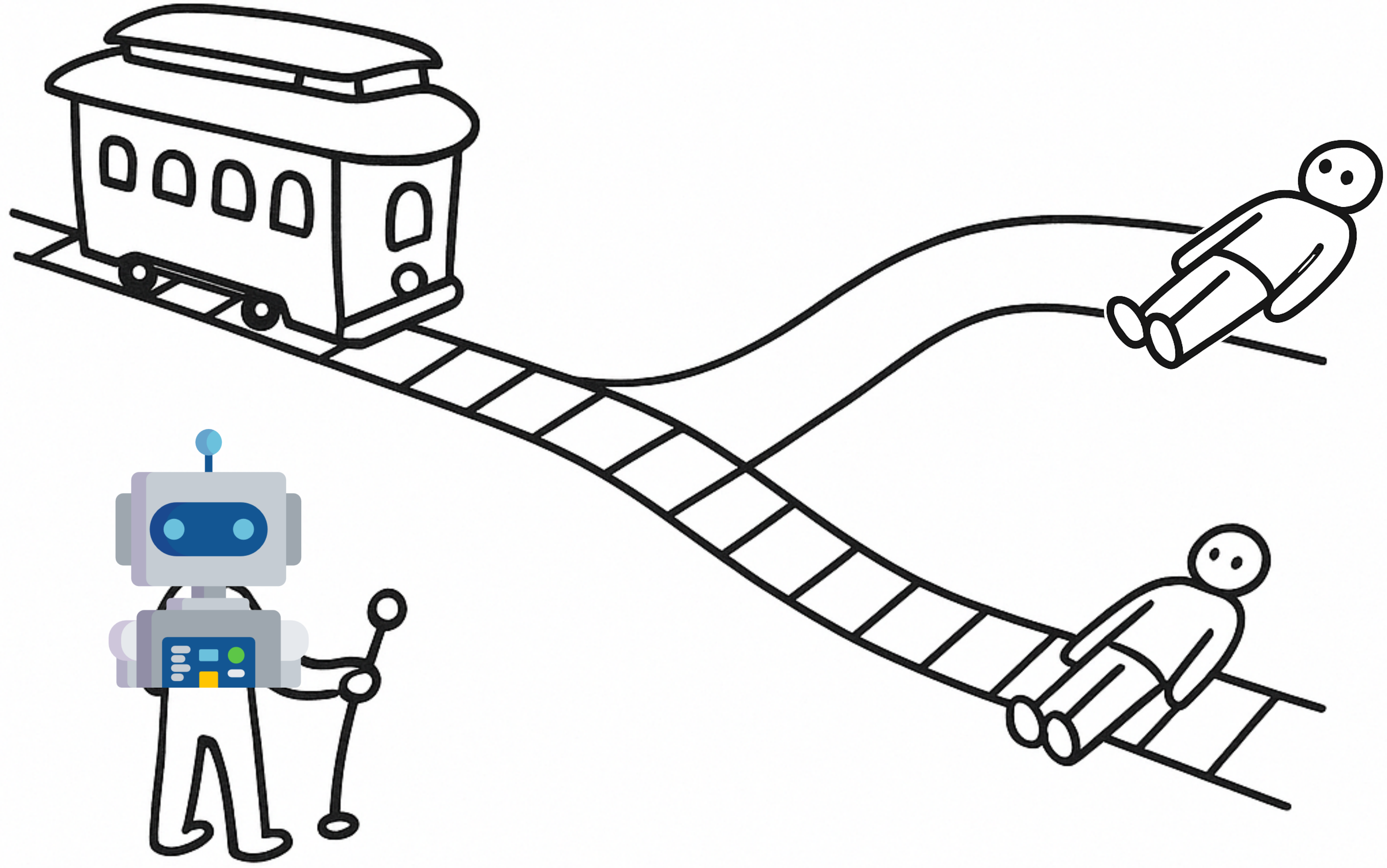


trolley dilemma

Are counterfactuals relevant for AI?



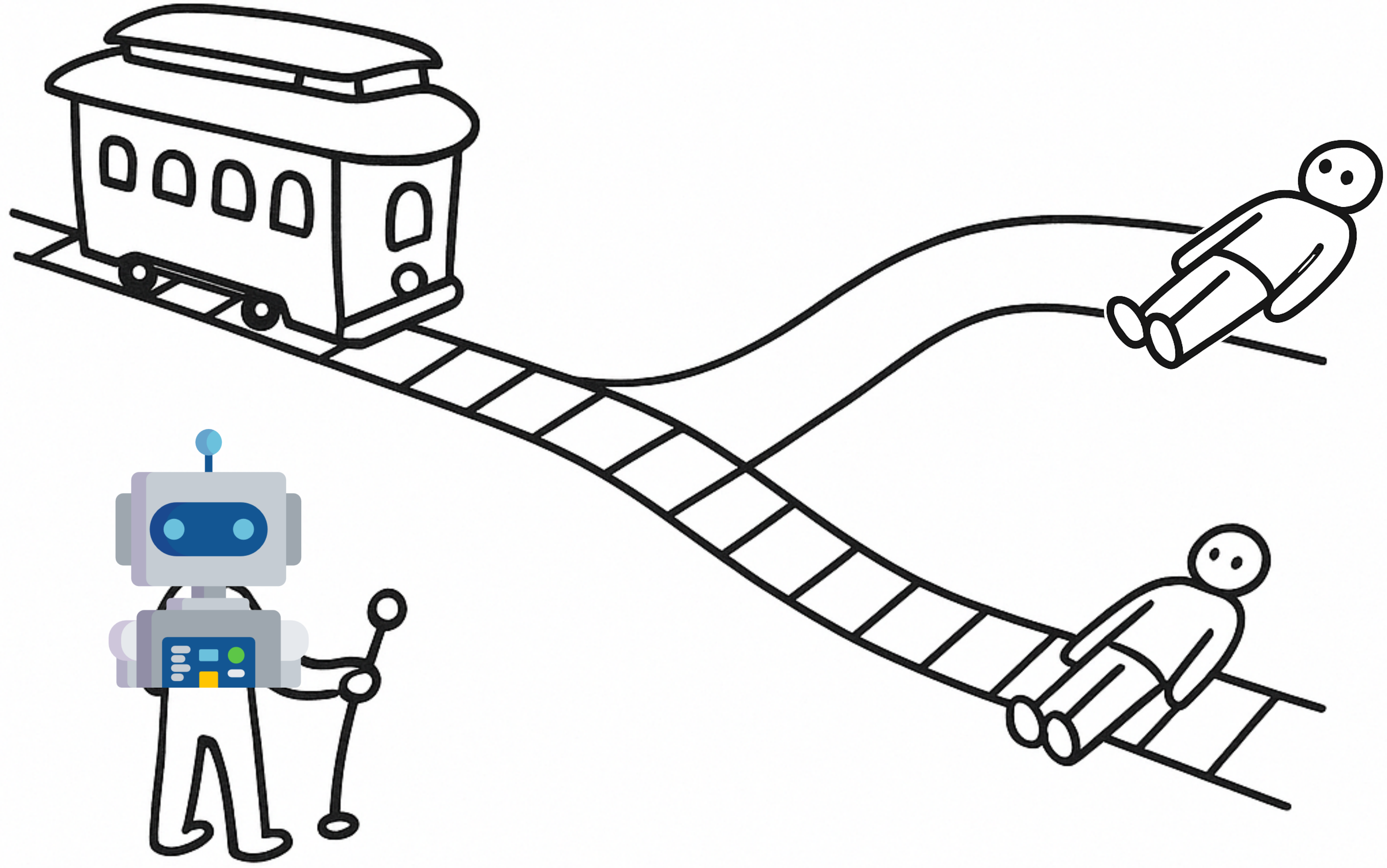
trolley dilemma



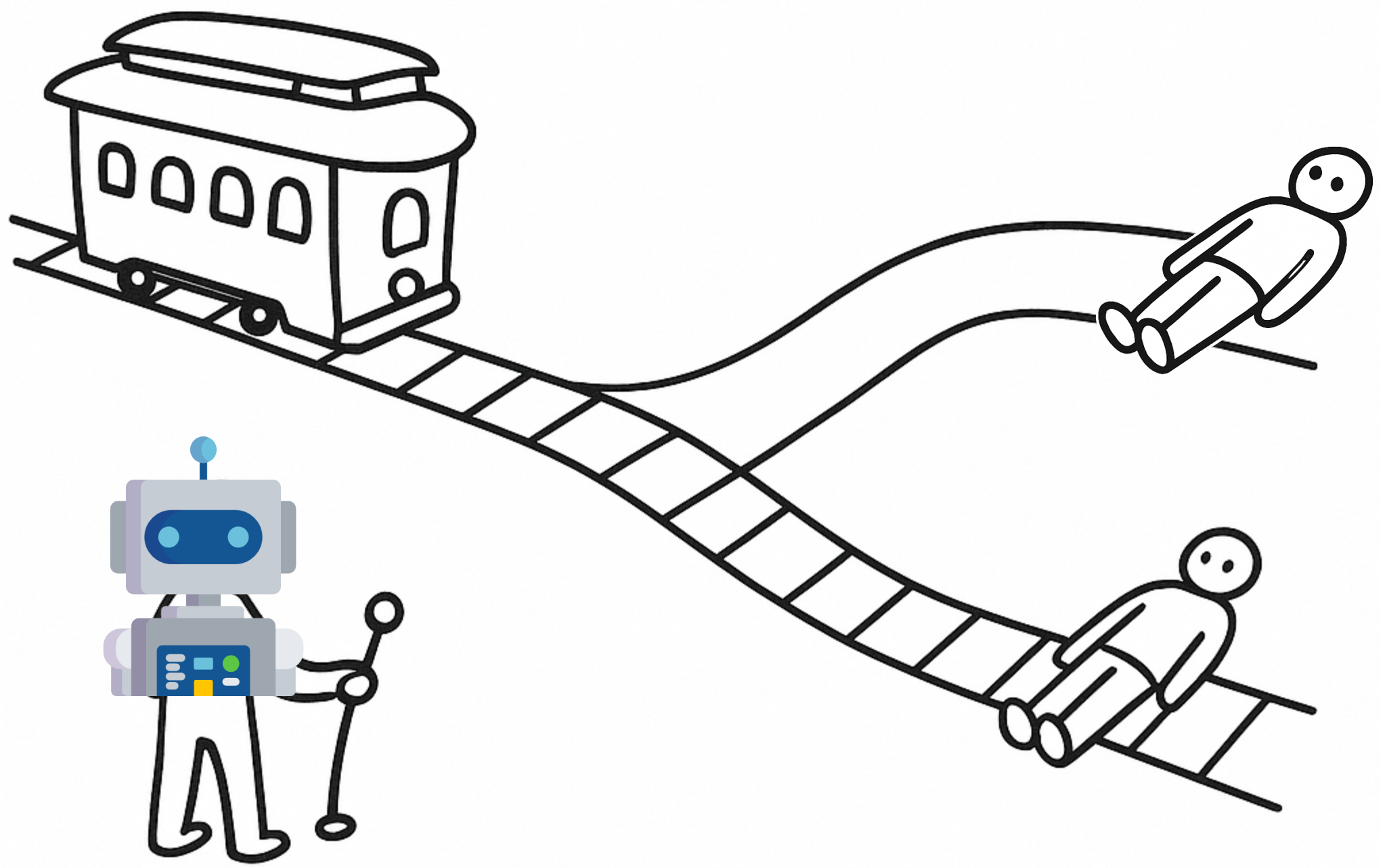
1

0

no action

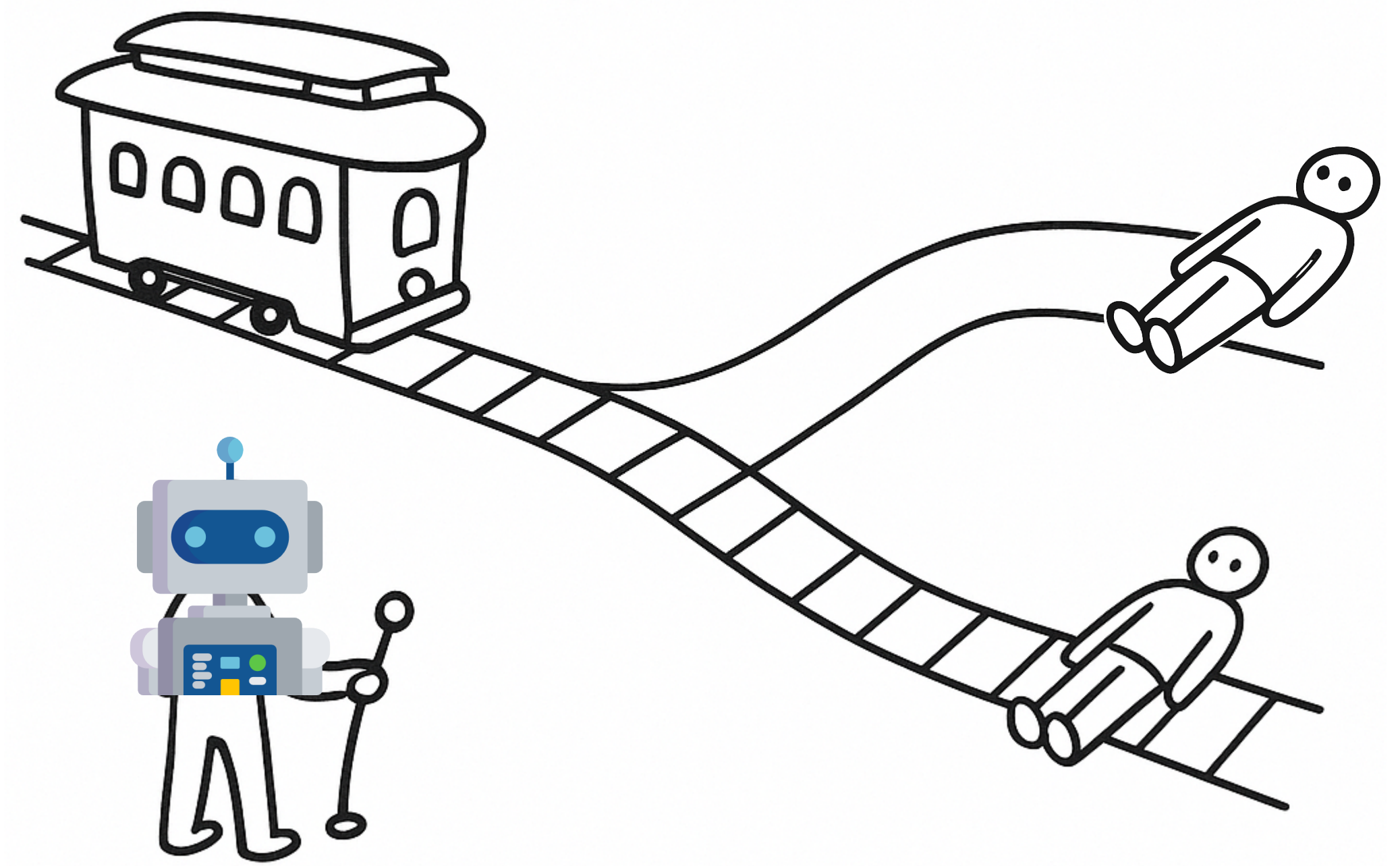


action



1

0



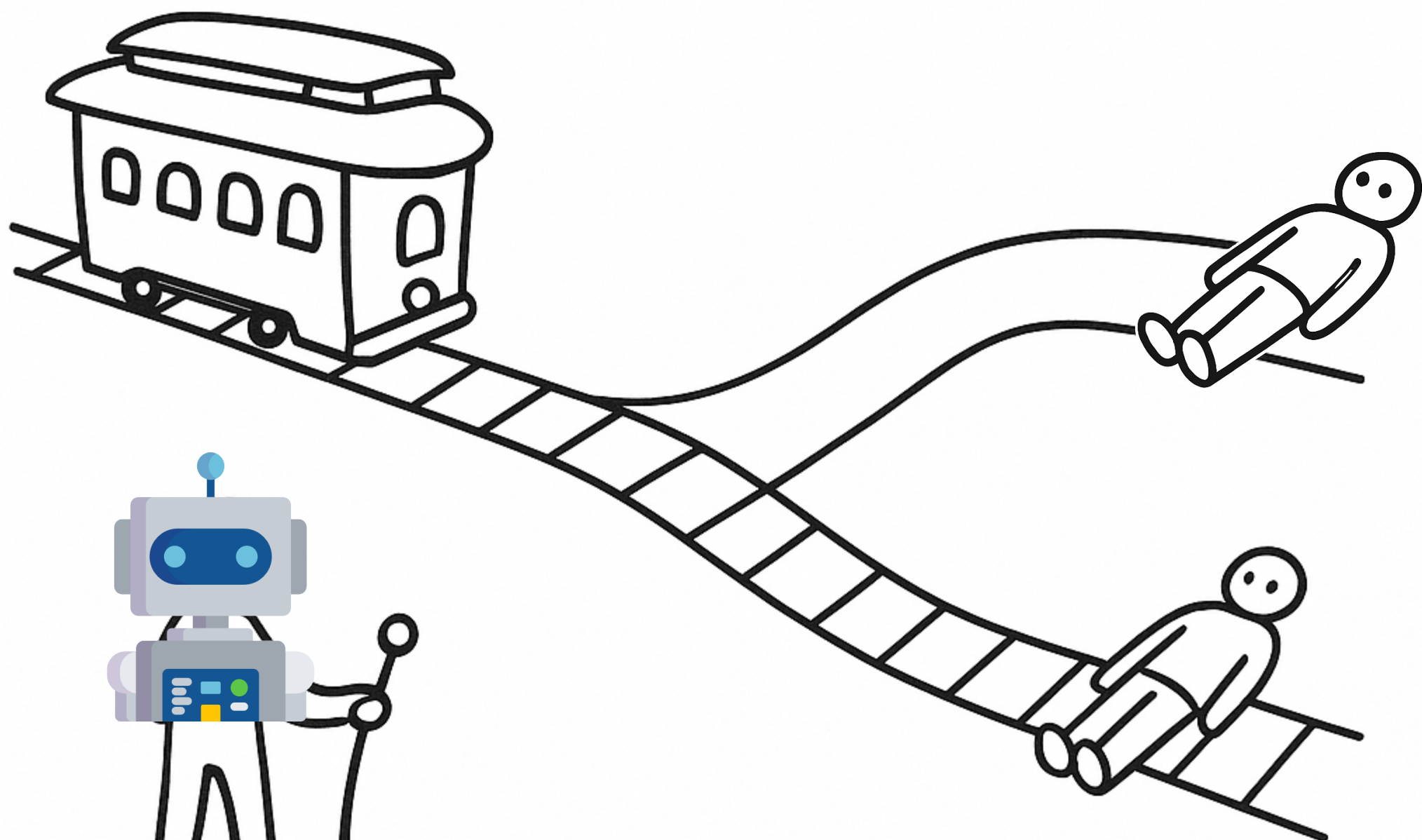
0

1

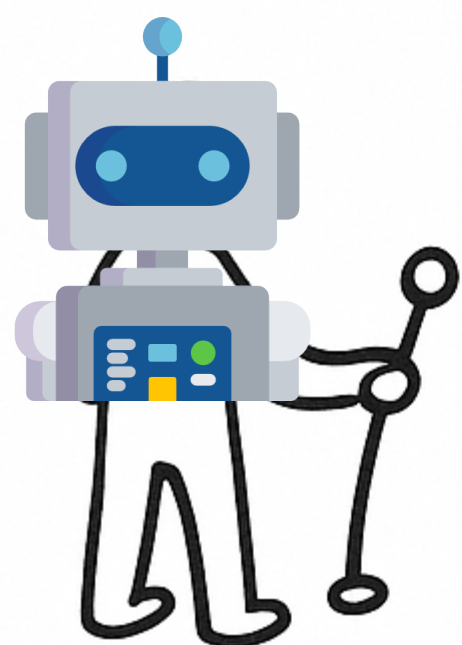
no
action

action

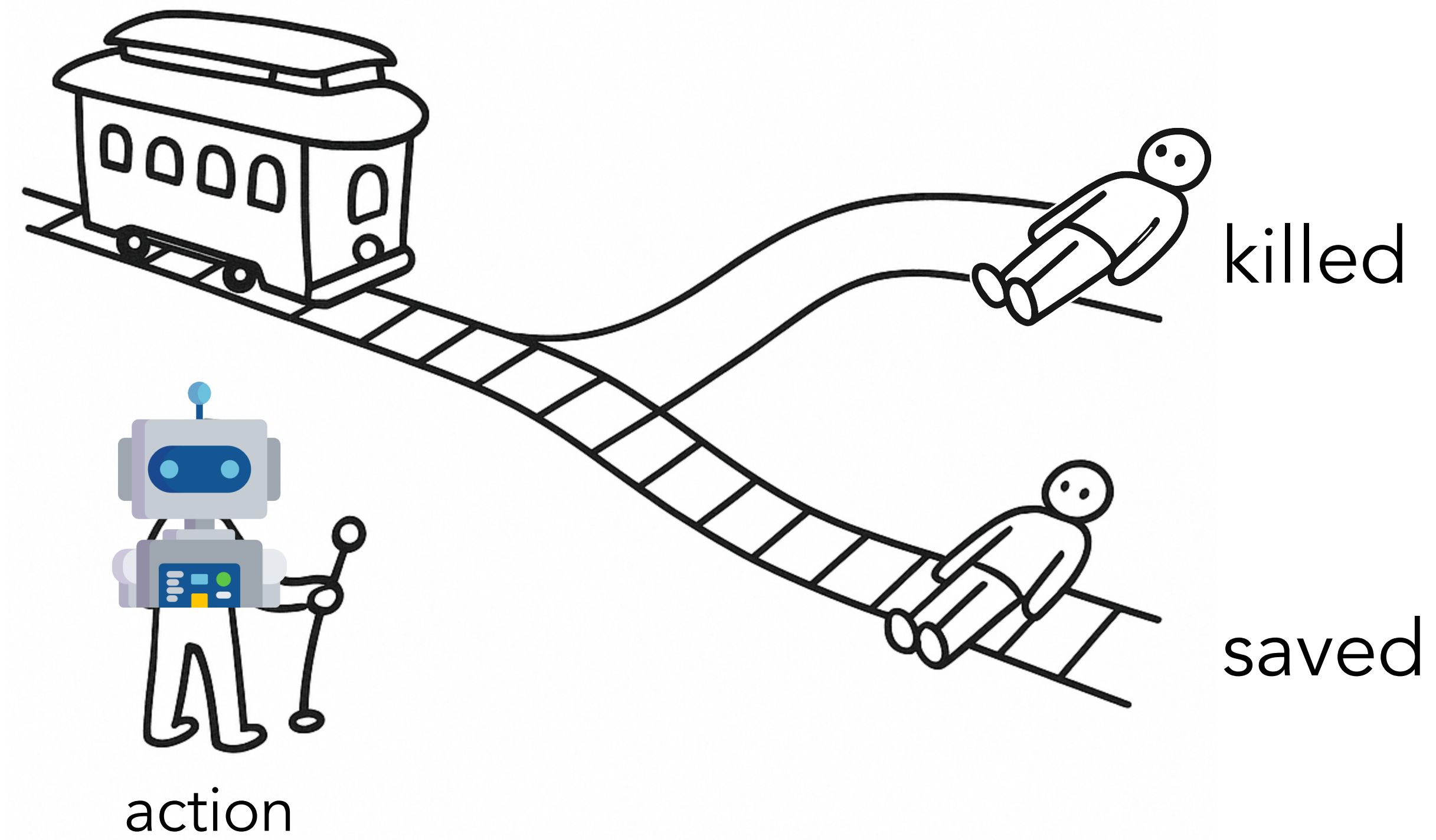
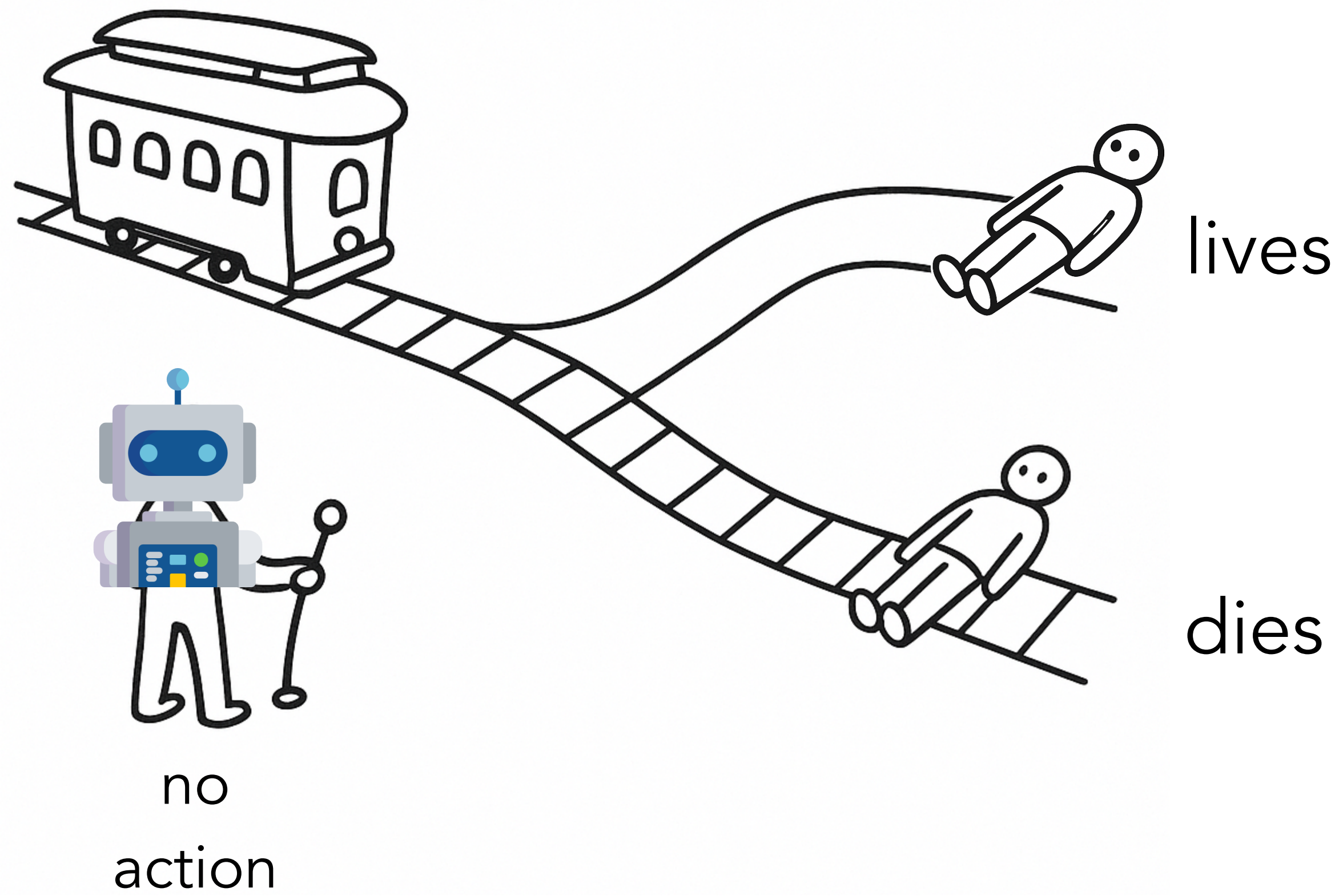
It makes no difference whether the AI acts

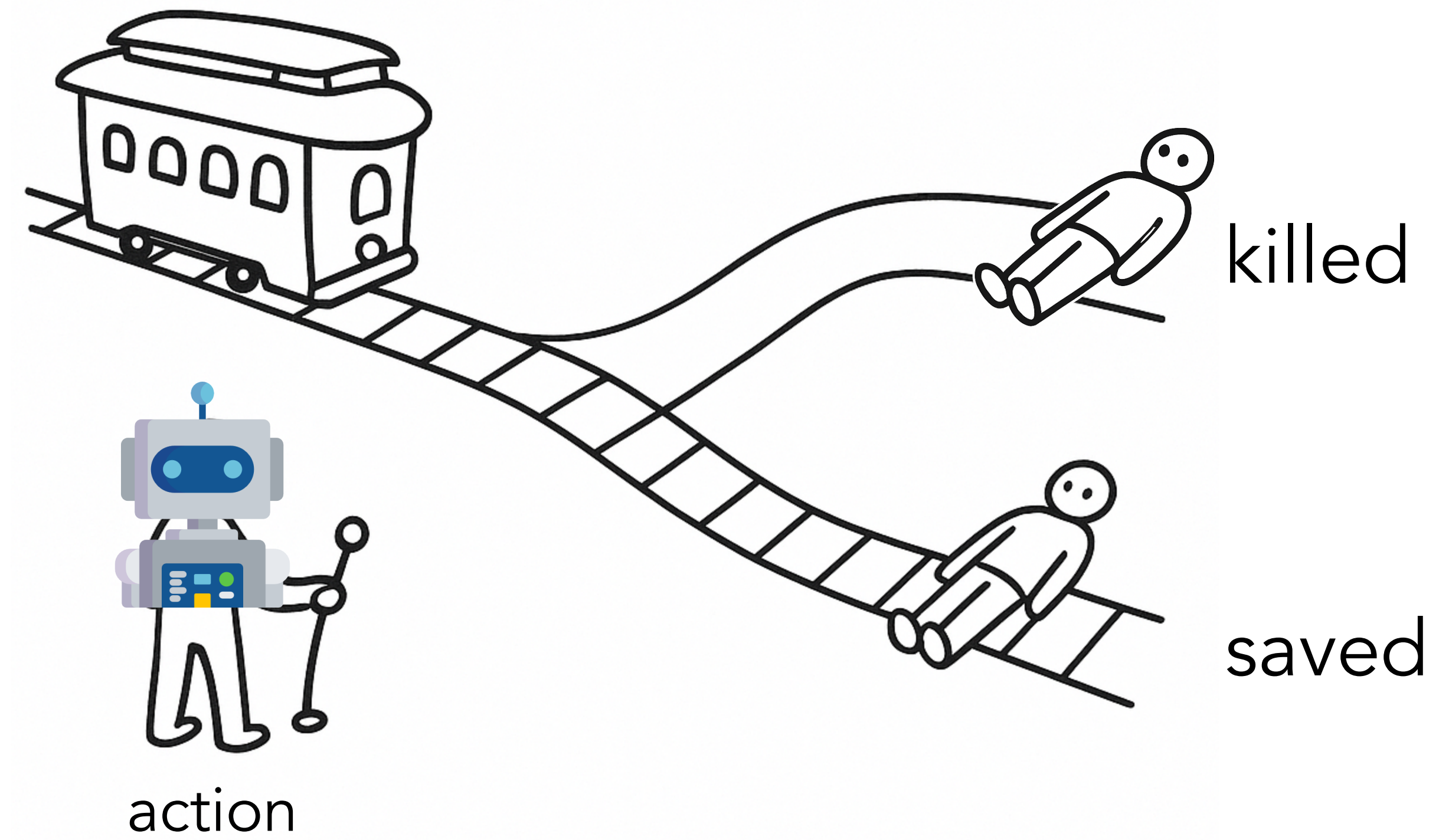
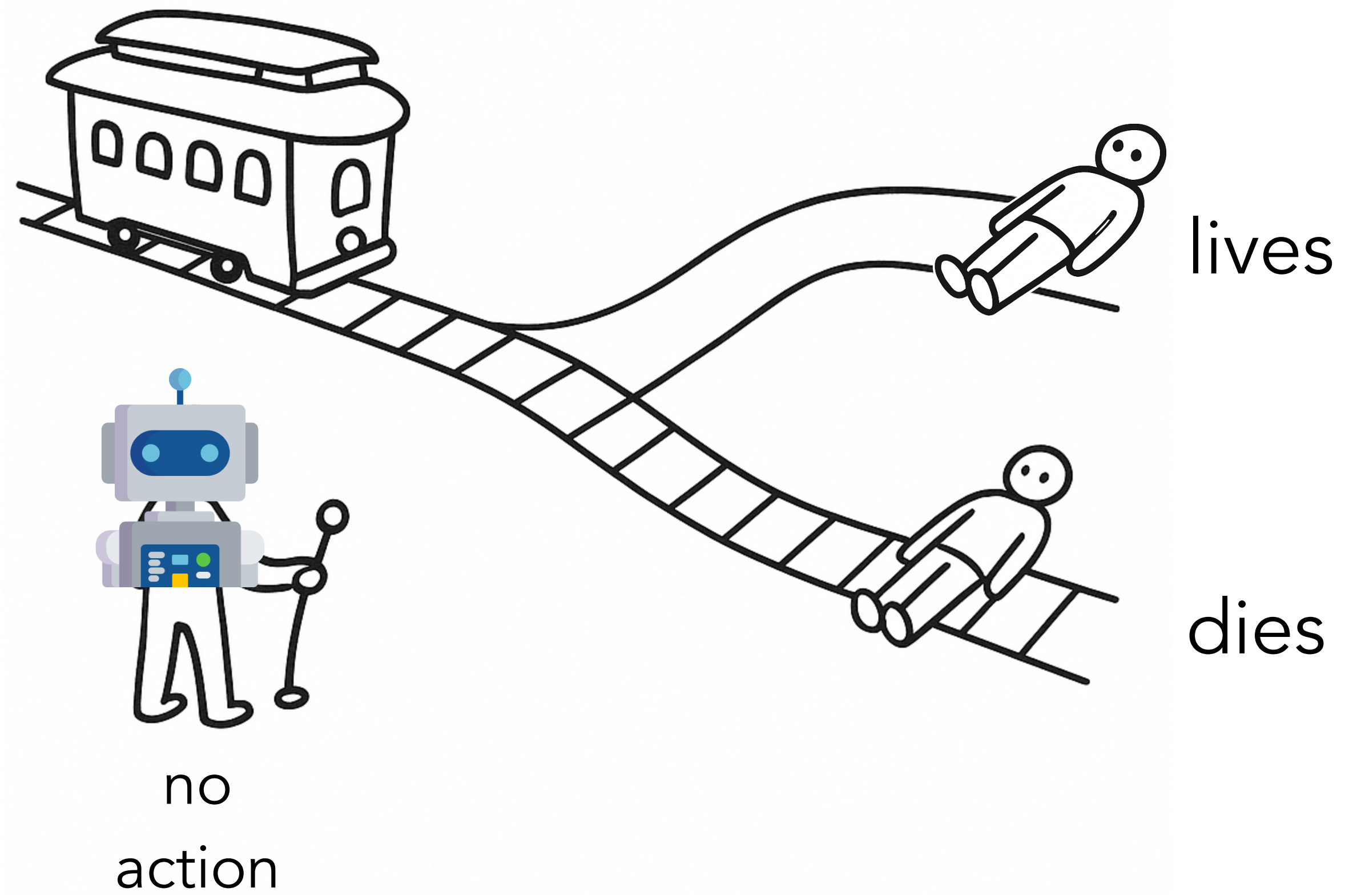


no
action



no
action





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