c i c Causality in Cognition

Beyond the here and now

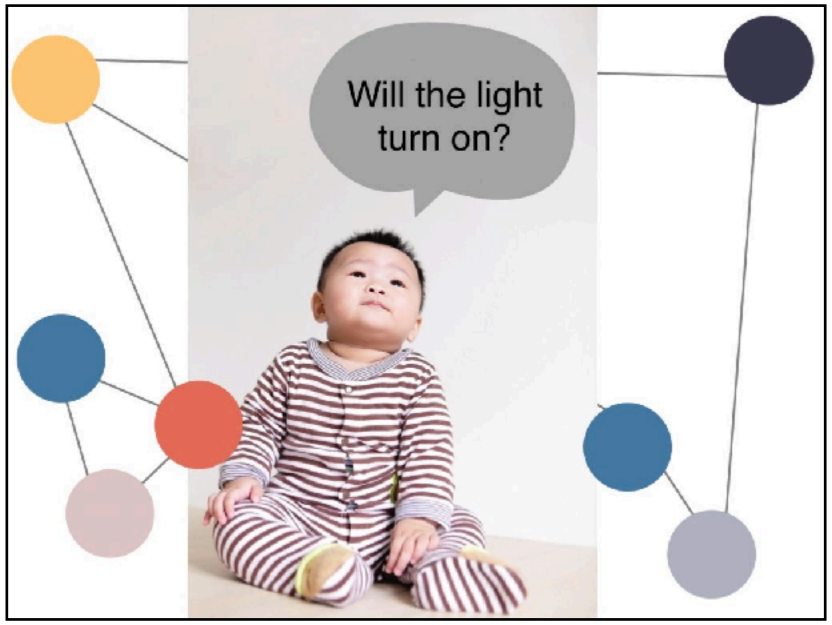
Counterfactual simulation in human cognition

@tobigerstenberg

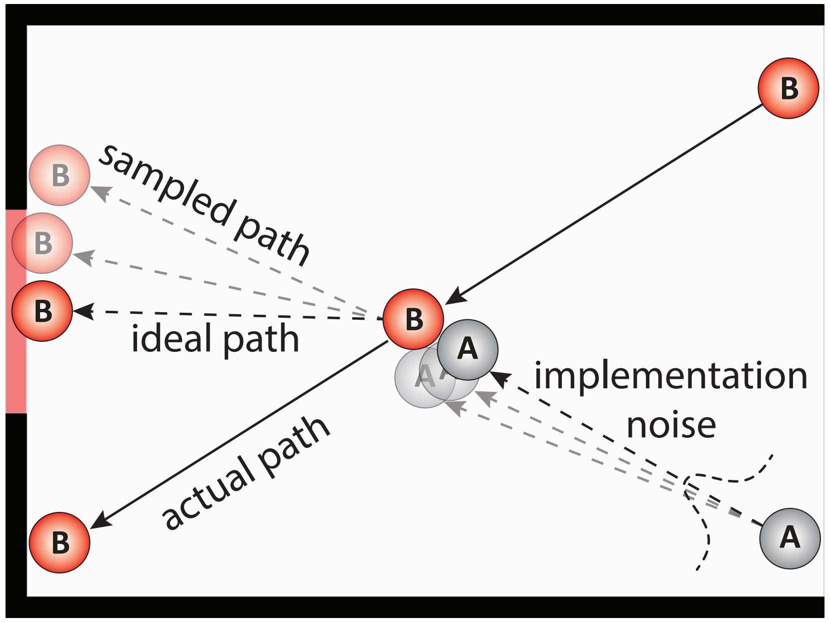
http://cicl.stanford.edu

Causality in Cognition

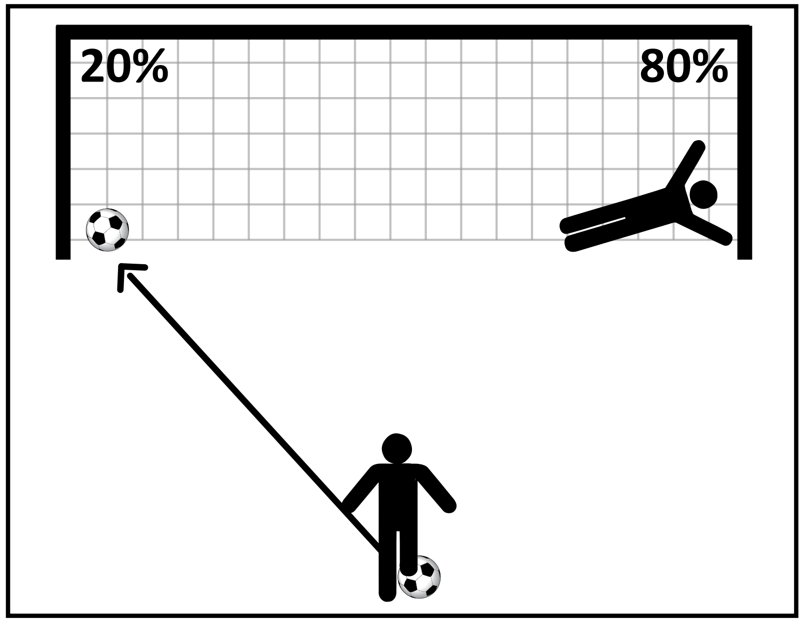
Our lab studies the role of causality in people's understanding of the world, and of each other.



learning



reasoning



judgment

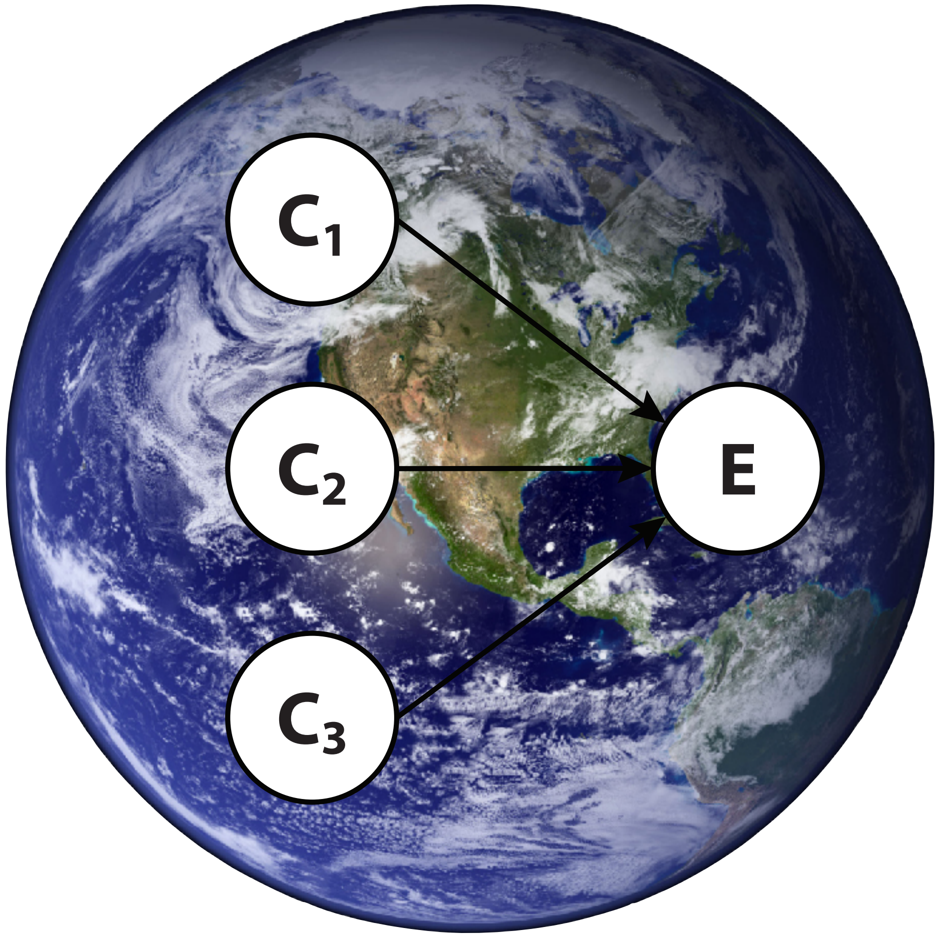
@tobigerstenberg

http://cicl.stanford.edu

A computational framework for understanding responsibility

What causal role did the action play?

What does the action reveal about the person?



Intuitive theory of how **the world** works

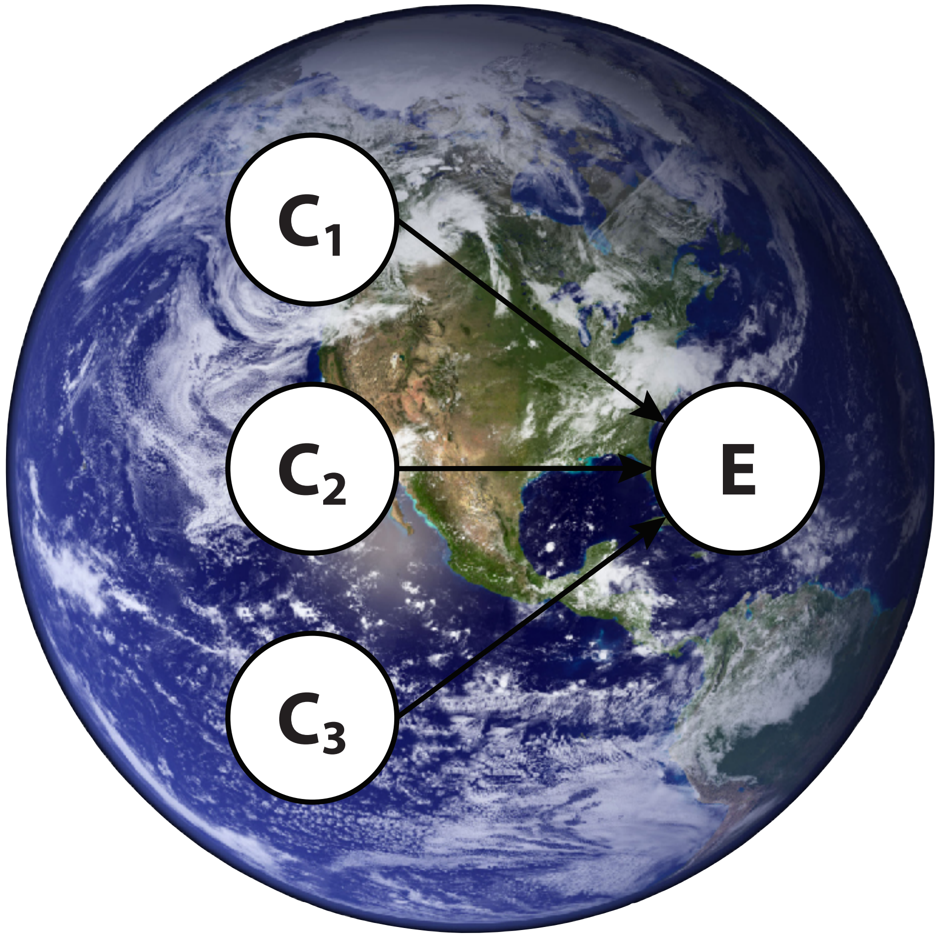
Intuitive theory of how **people** work

Gerstenberg & Tenenbaum (2017) Intuitive Theories. *Oxford Handbook of Causal Reasoning*

Goodman, Tenenbaum, & Gerstenberg (2015) Concepts in a probabilistic language of thought. *The Conceptual Mind: New Directions in the Study of Concepts*

A computational framework for understanding responsibility

What causal role
did the action play?

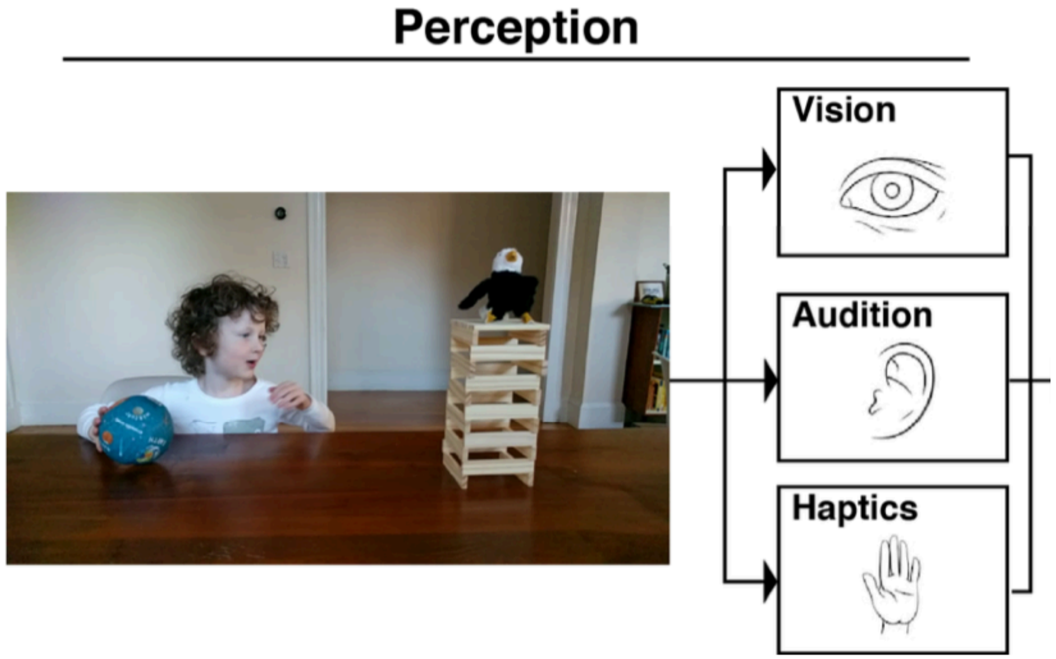


Intuitive theory of
how **the world** works

Gerstenberg & Tenenbaum (2017) Intuitive Theories. *Oxford Handbook of Causal Reasoning*

Goodman, Tenenbaum, & Gerstenberg (2015) Concepts in a probabilistic language of thought. *The Conceptual Mind: New Directions in the Study of Concepts*

Mental models: The **physics engine** in the head



infer
the past

explain
the present

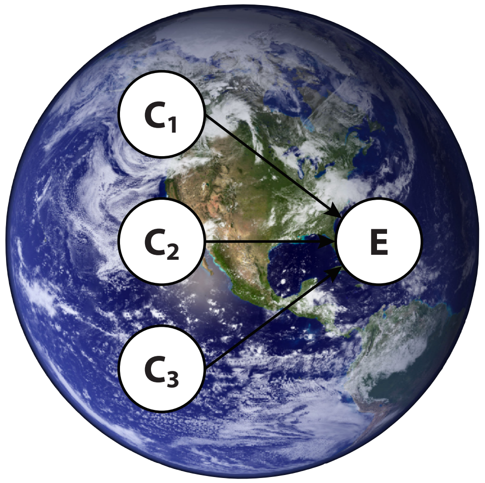
predict
the future

Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (in press). Probabilistic models of physical reasoning. In T. L. Griffiths, N. Chater, & J. B. Tenenbaum (Eds.), *Reverse engineering the mind: Probabilistic models of cognition.*

When we want to **explain what happened** and **why**, we have to go beyond the here and now.

3 key ingredients for giving causal explanations

Mental models



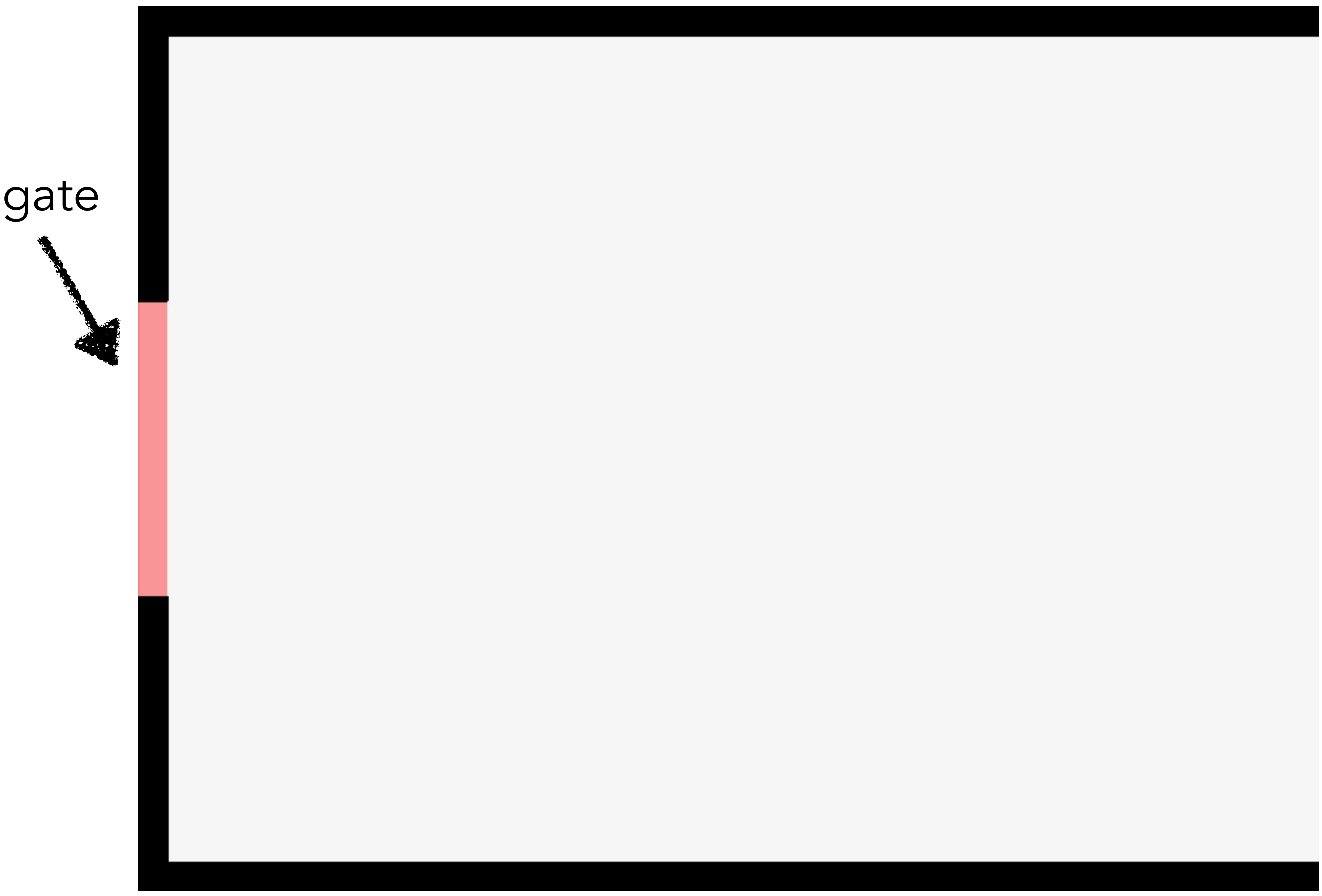
Counterfactual interventions

Mental simulation

How do people make **causal judgments** about physical events?

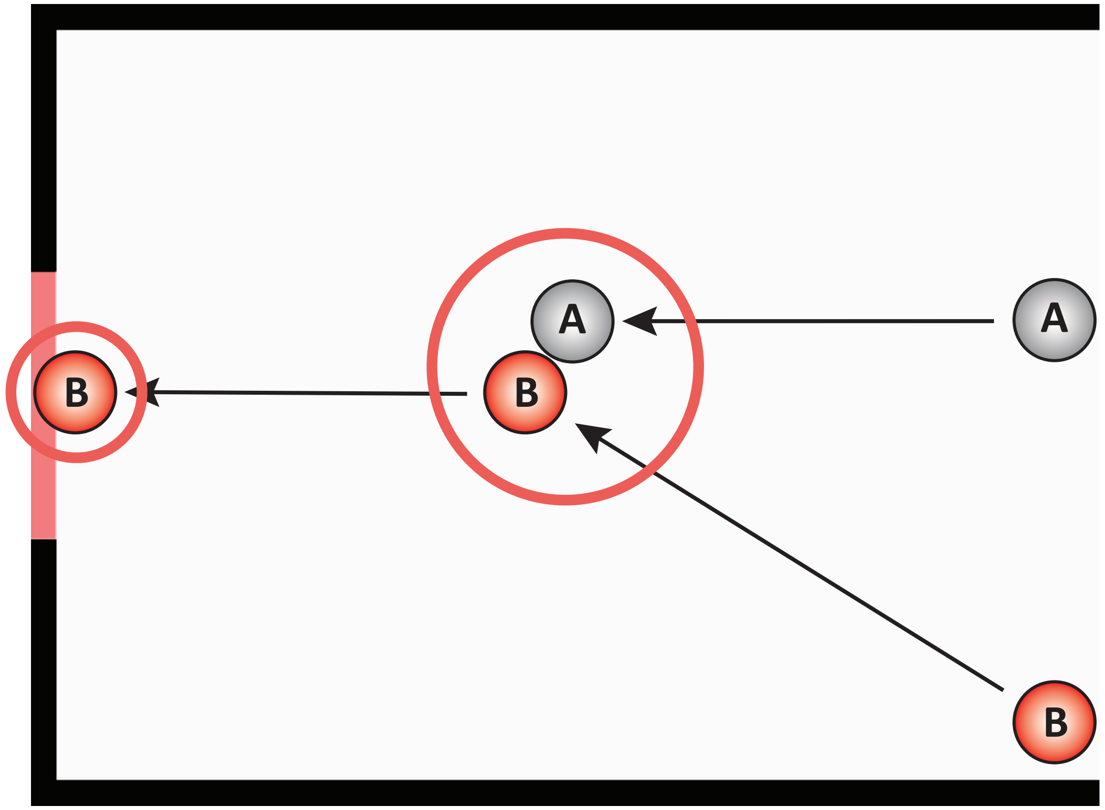
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Did A cause B to go through the gate?



Counterfactual Simulation Model

What happened?

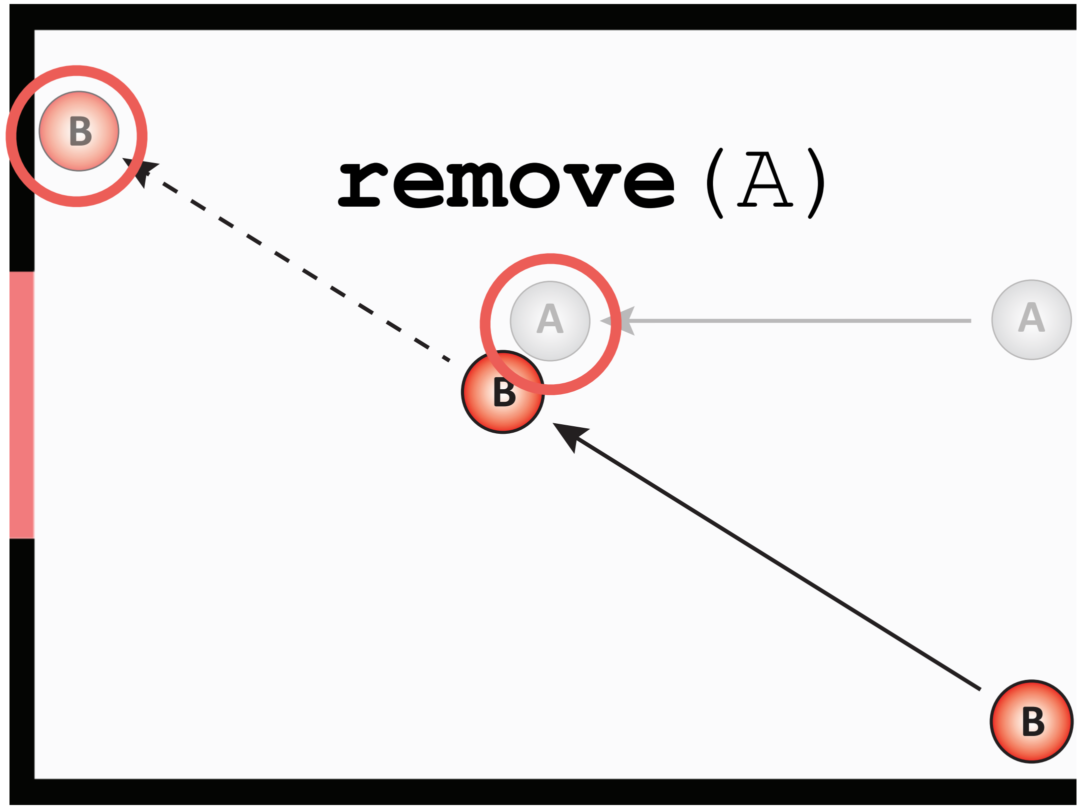


Actual situation

B went through the gate

≠ ?

What would have happened?



Counterfactual situation

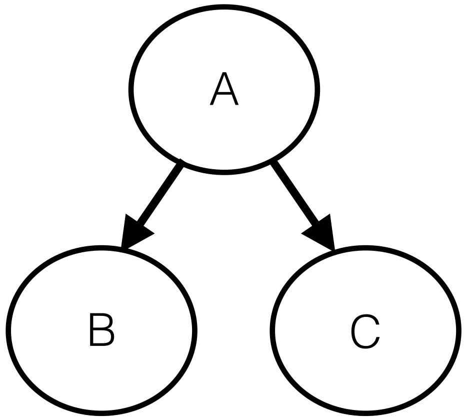
B would have missed the gate

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Causal judgments as counterfactual contrasts over generative models

Generative model

causal Bayes net



structural equations

$$B = A$$

$$C = A$$

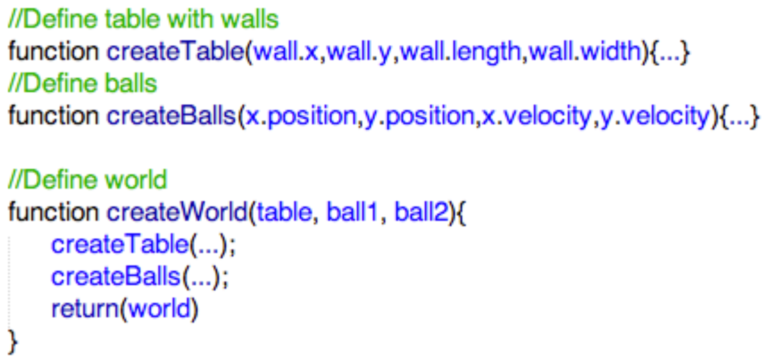
Counterfactual intervention

`do``()` operator

Pearl, J. (2000). *Causality: Models, reasoning and inference*

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)
}
```

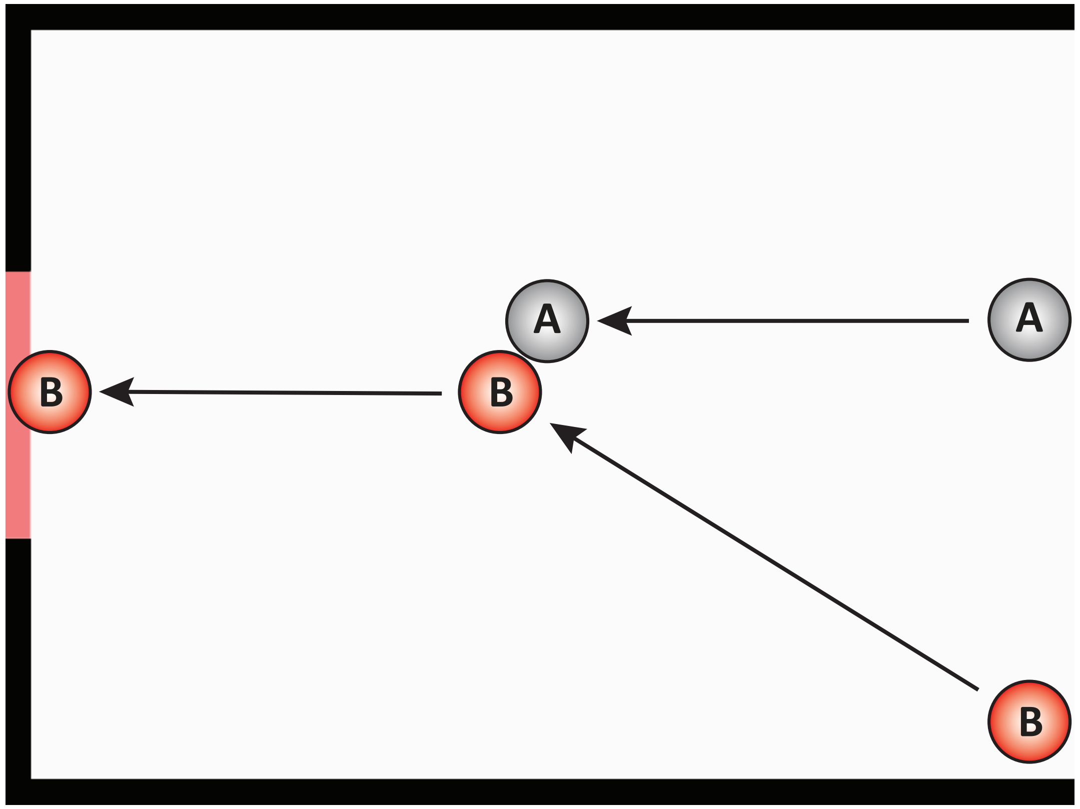
Counterfactual intervention

`remove``(object)` operator

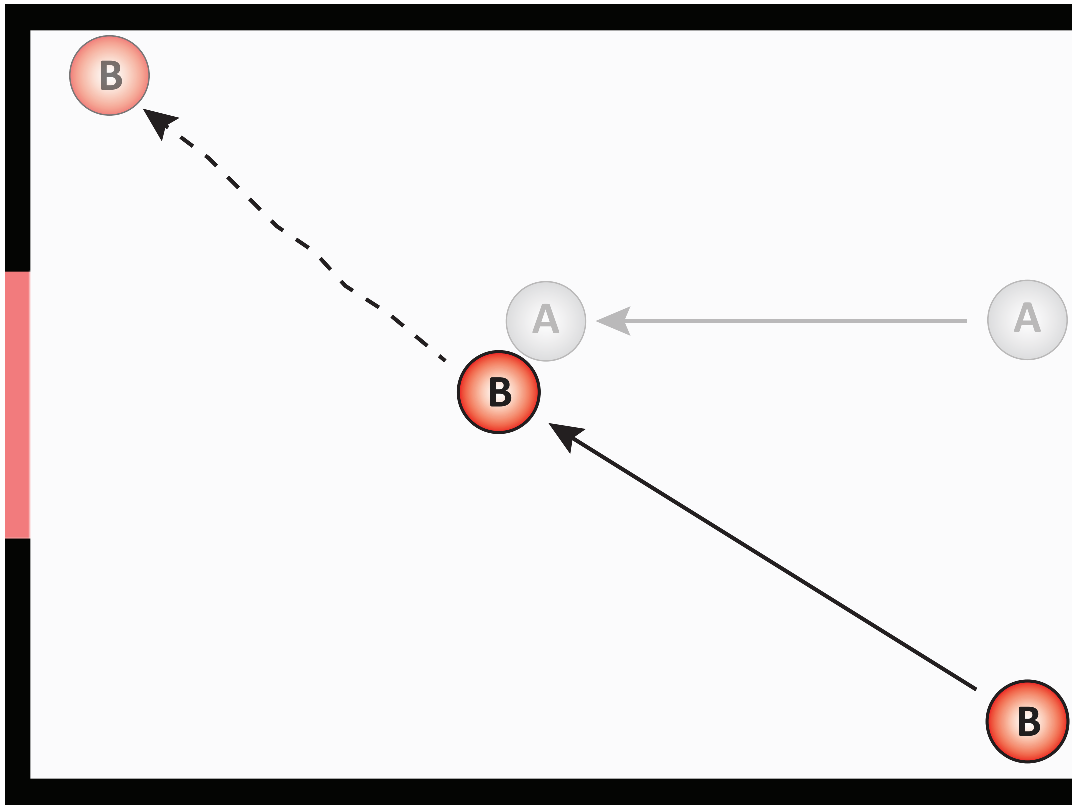
Chater & Oaksford (2013) Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*

Goodman, Tenenbaum, & Gerstenberg (2015) Concepts in a probabilistic language of thought. *The Conceptual Mind: New Directions in the Study of Concepts*

What happened?



What would have happened?



Actual situation

B went through the gate

$\overset{?}{\neq}$

Counterfactual situation

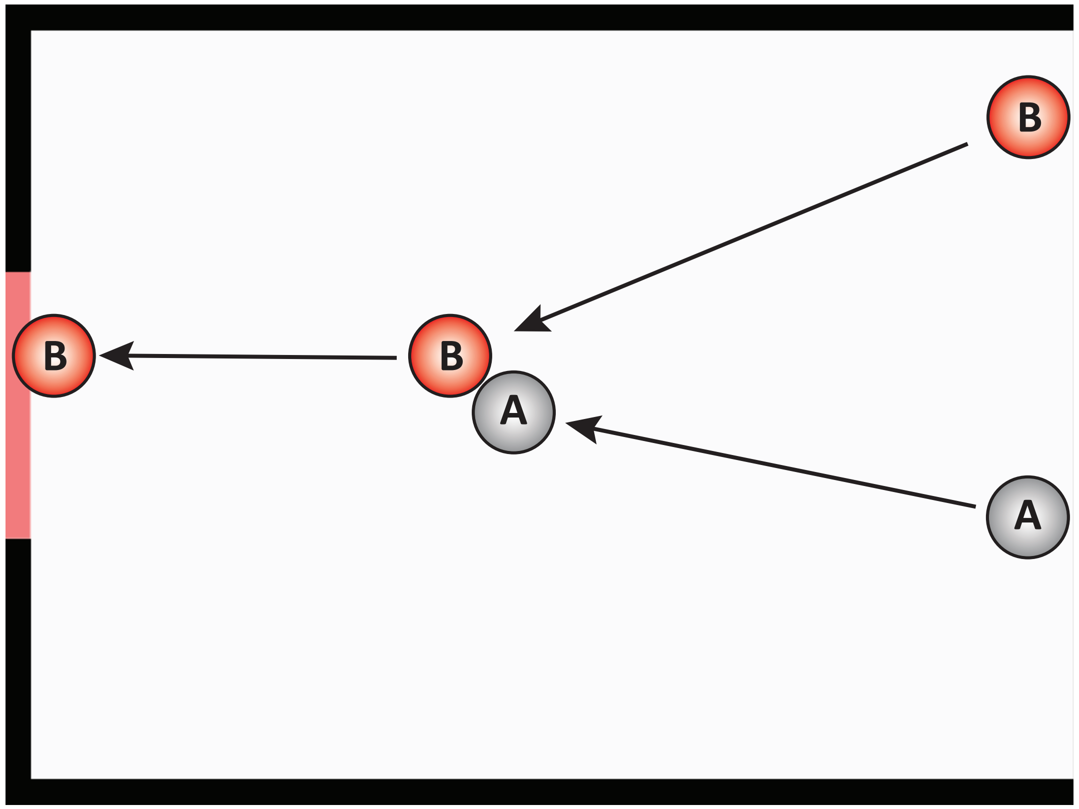
B would have missed the gate ✓

B would have missed the gate ✓

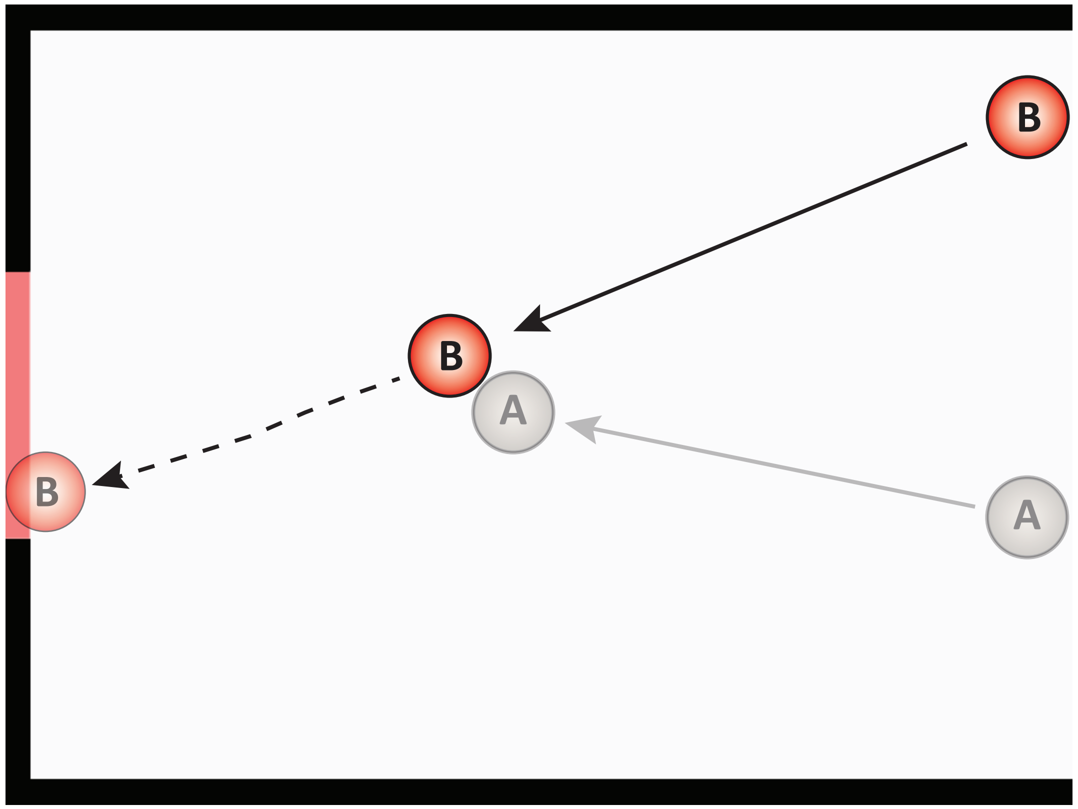
B would have missed the gate ✓

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

What happened?



What would have happened?



Actual situation

B went through the gate

?≠

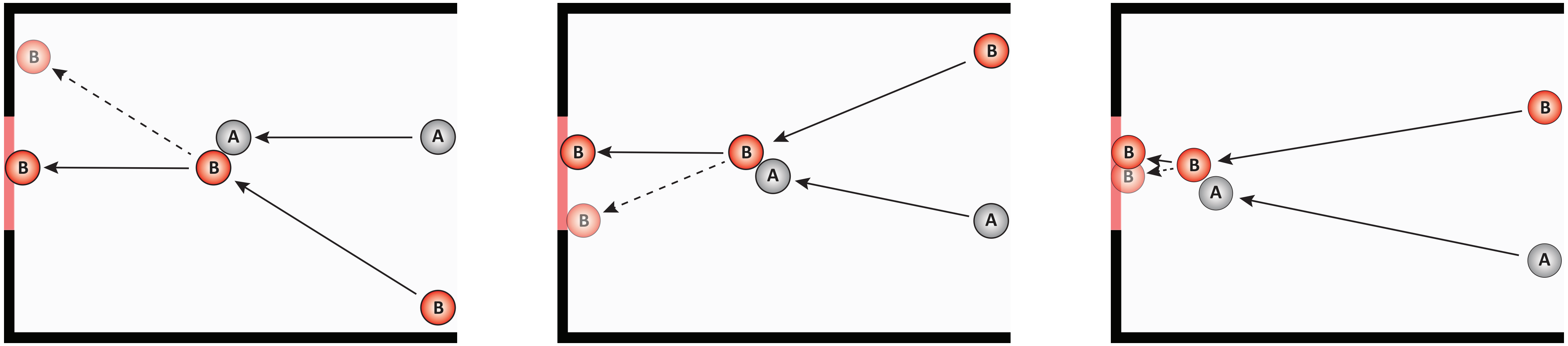
Counterfactual situation

B would have missed the gate ✓

B would have gone through gate ✗

B would have gone through gate ✗

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

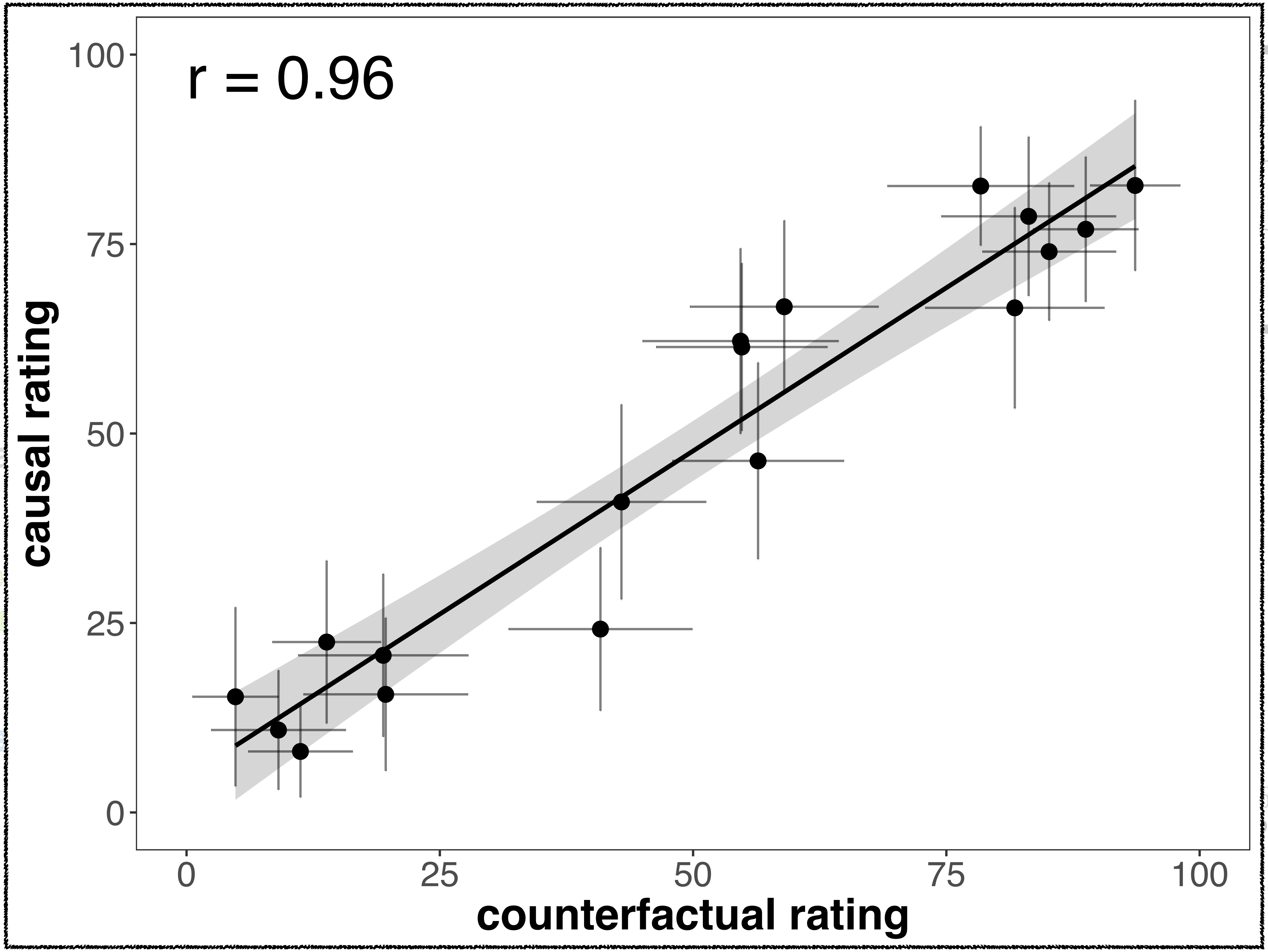


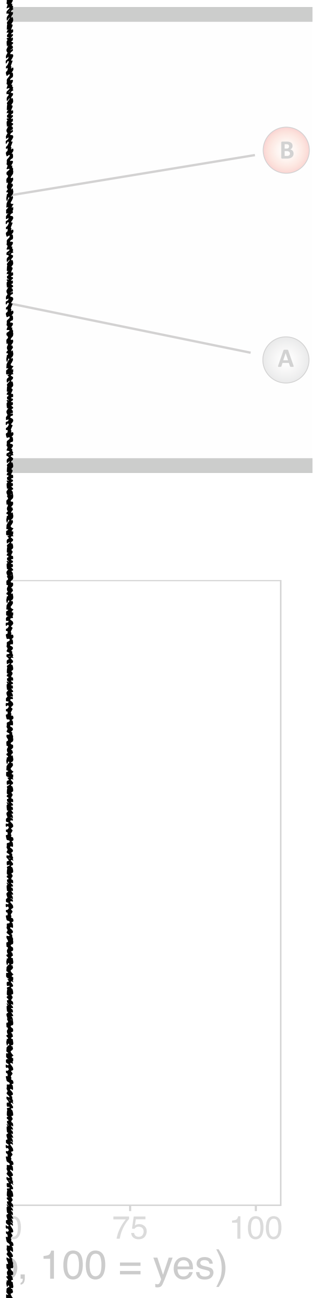
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Counterfa
simulation

Did A cause
through the

Would B
missed the



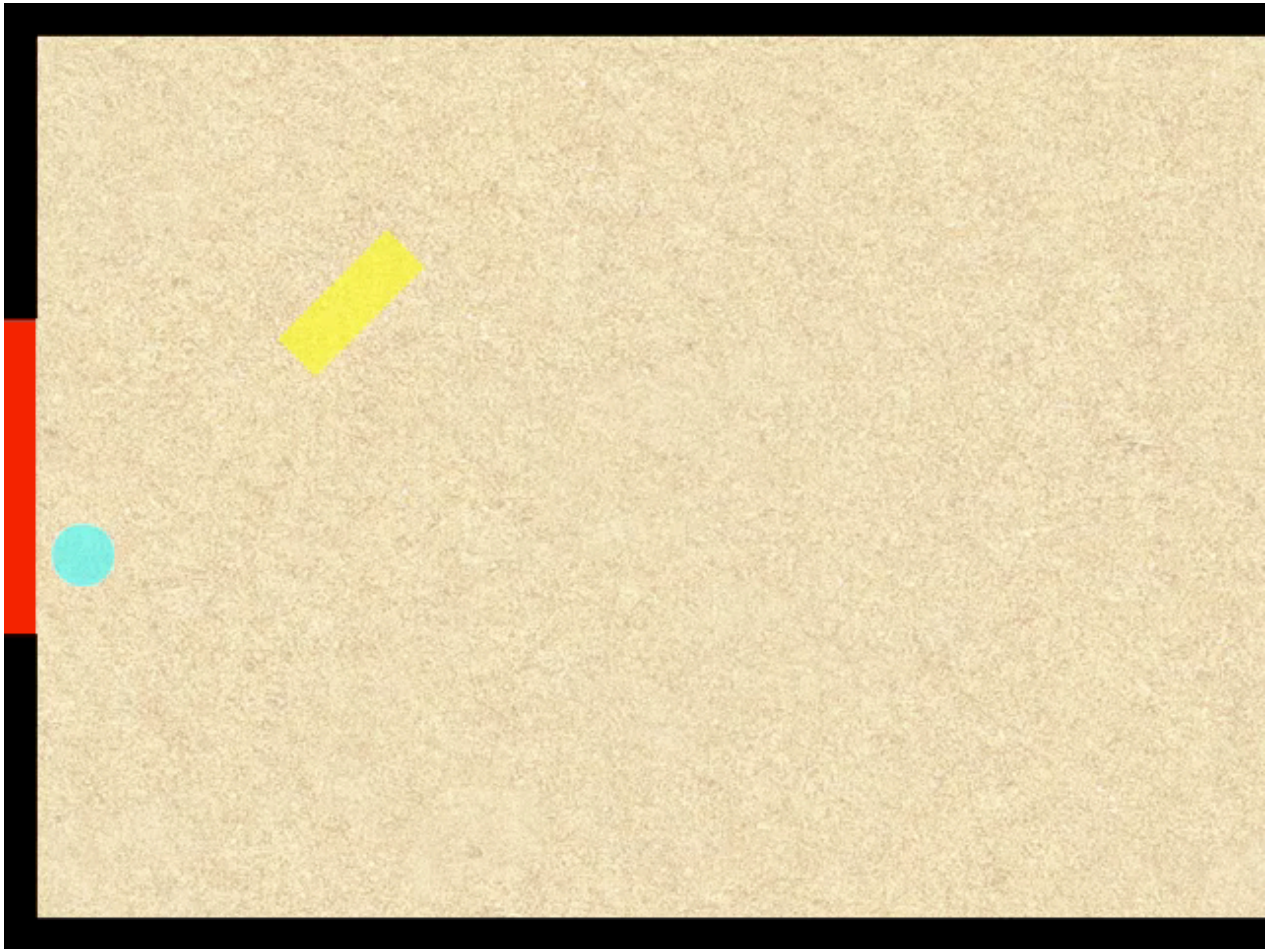


Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

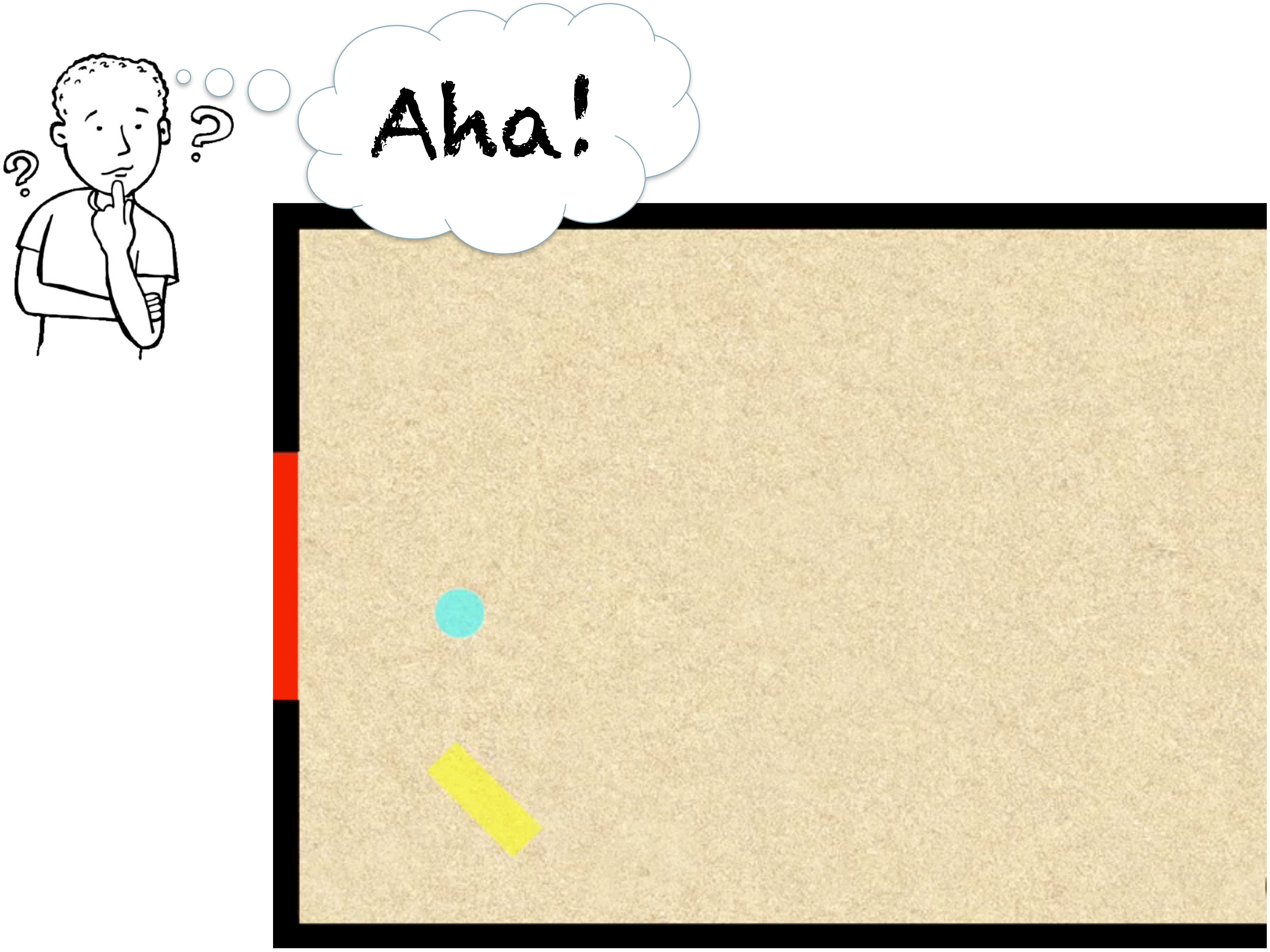
Are counterfactuals **necessary** for understanding causal judgments?

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Did A prevent B from going through the gate?



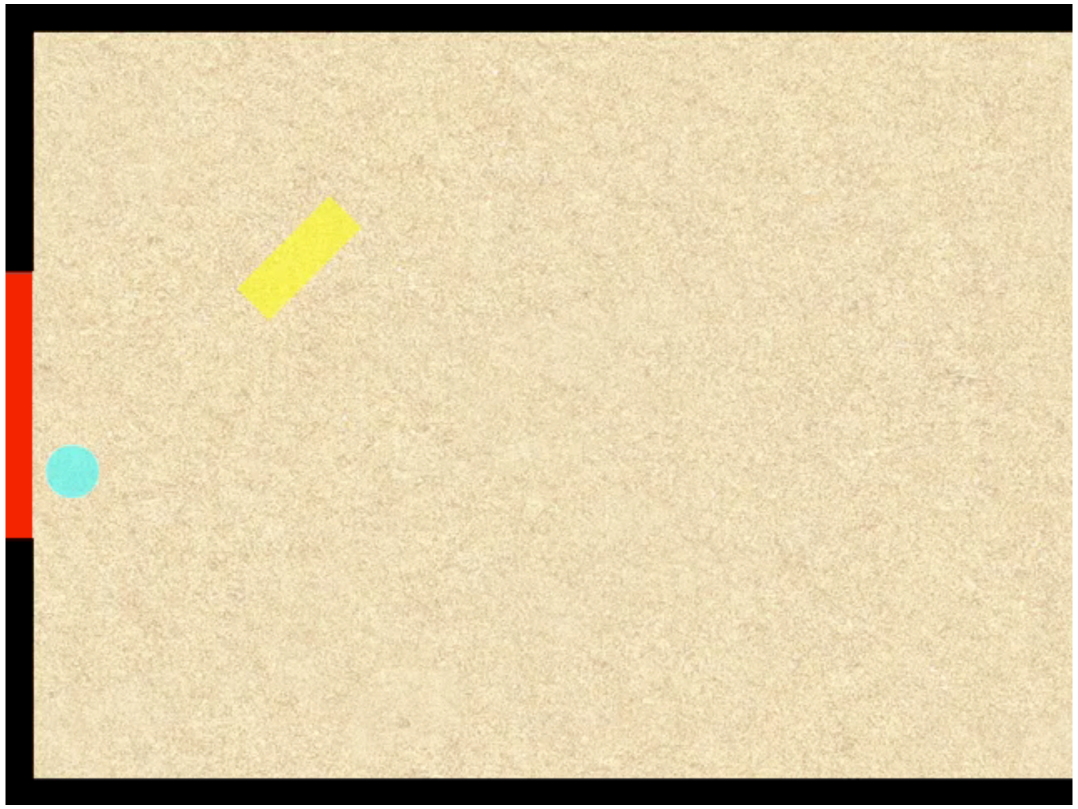
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*



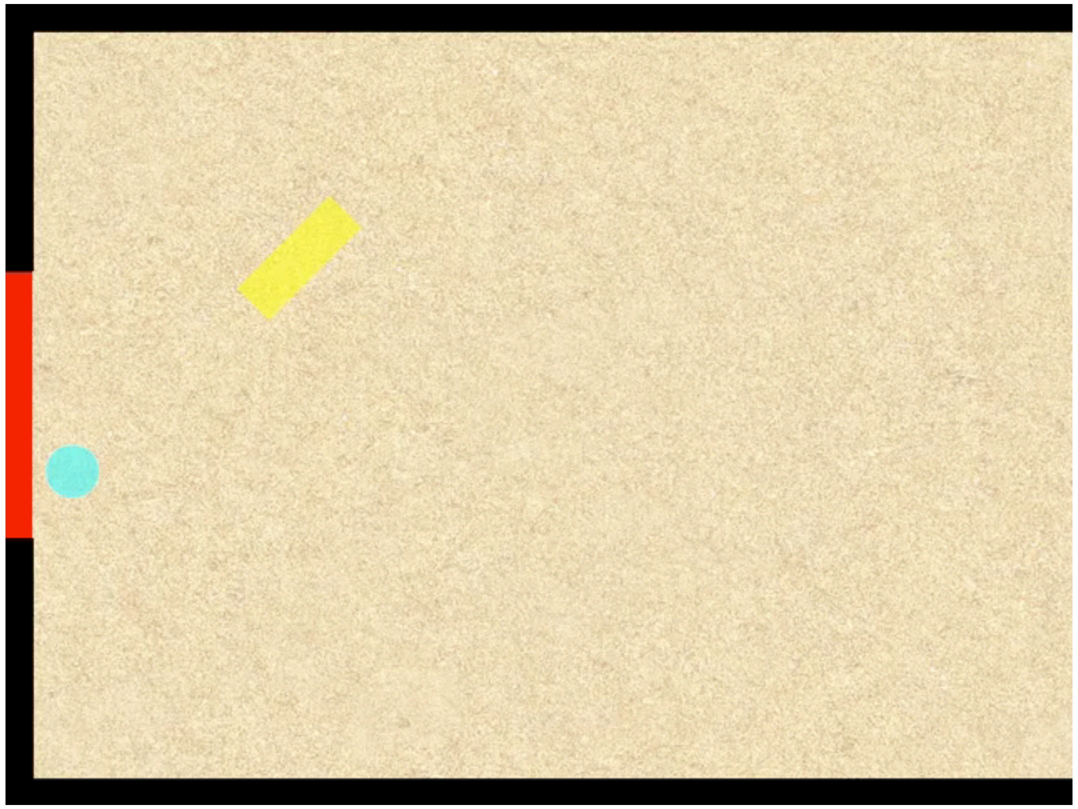
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Did A prevent B from going through the gate?

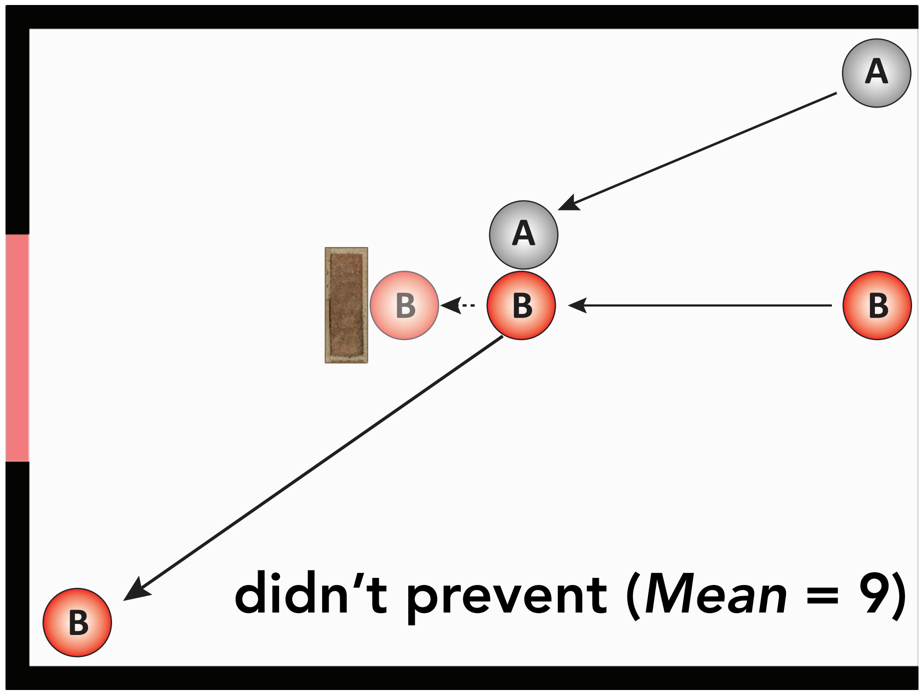
Actual



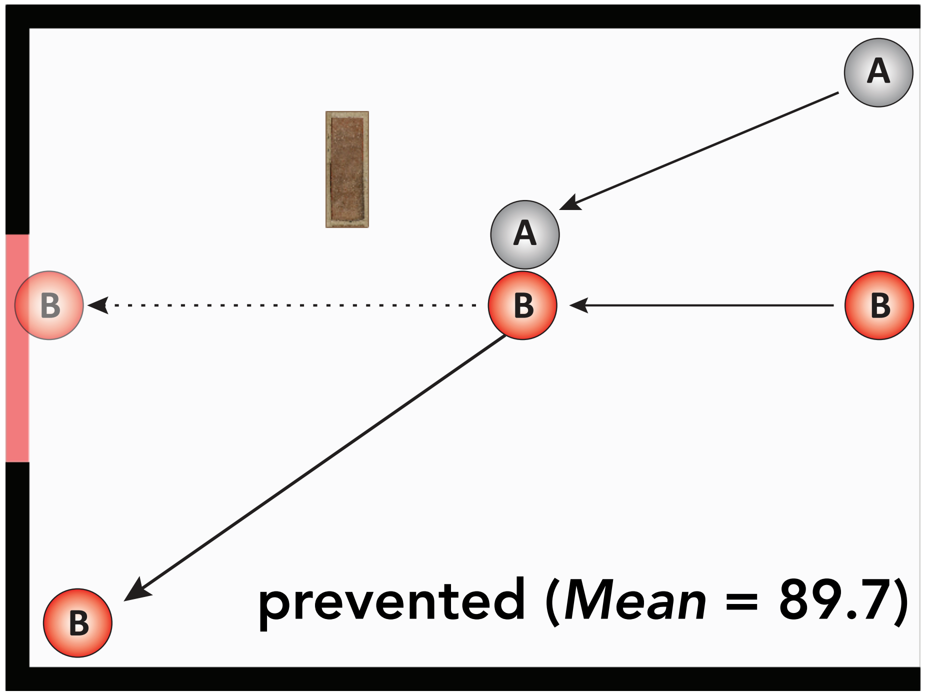
Counterfactual



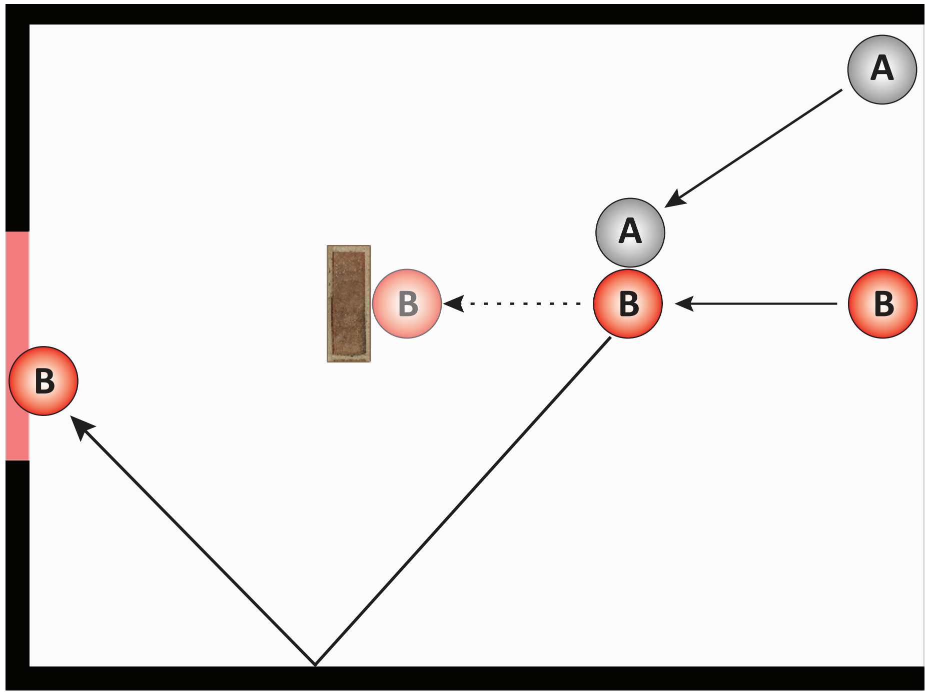
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*



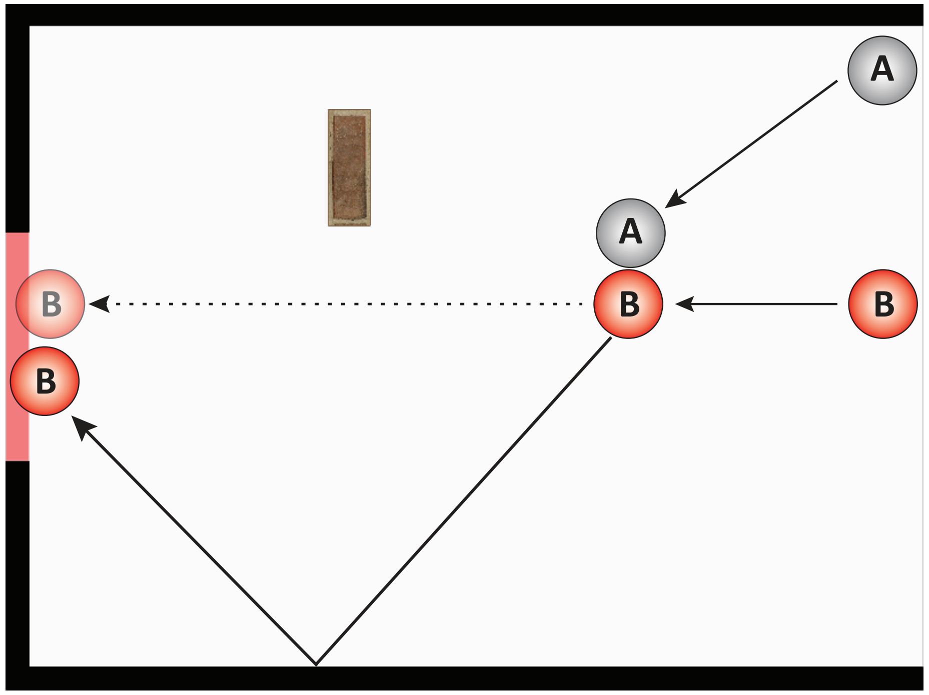
didn’t prevent (*Mean* = 9)



prevented (*Mean* = 89.7)



caused (*Mean* = 86.7)

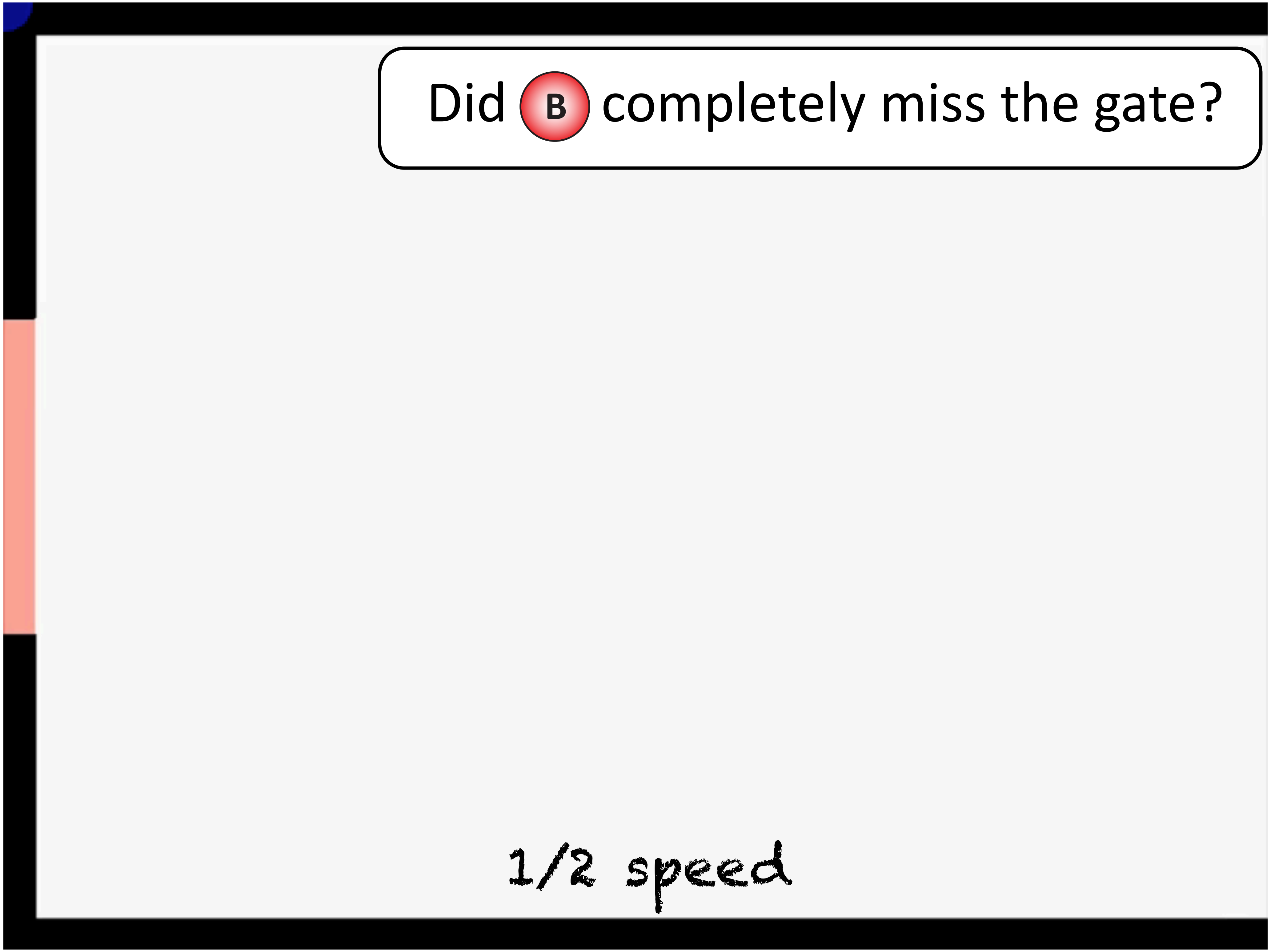


didn’t cause (*Mean* = 18.6)

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

How do people make causal judgments about physical events?

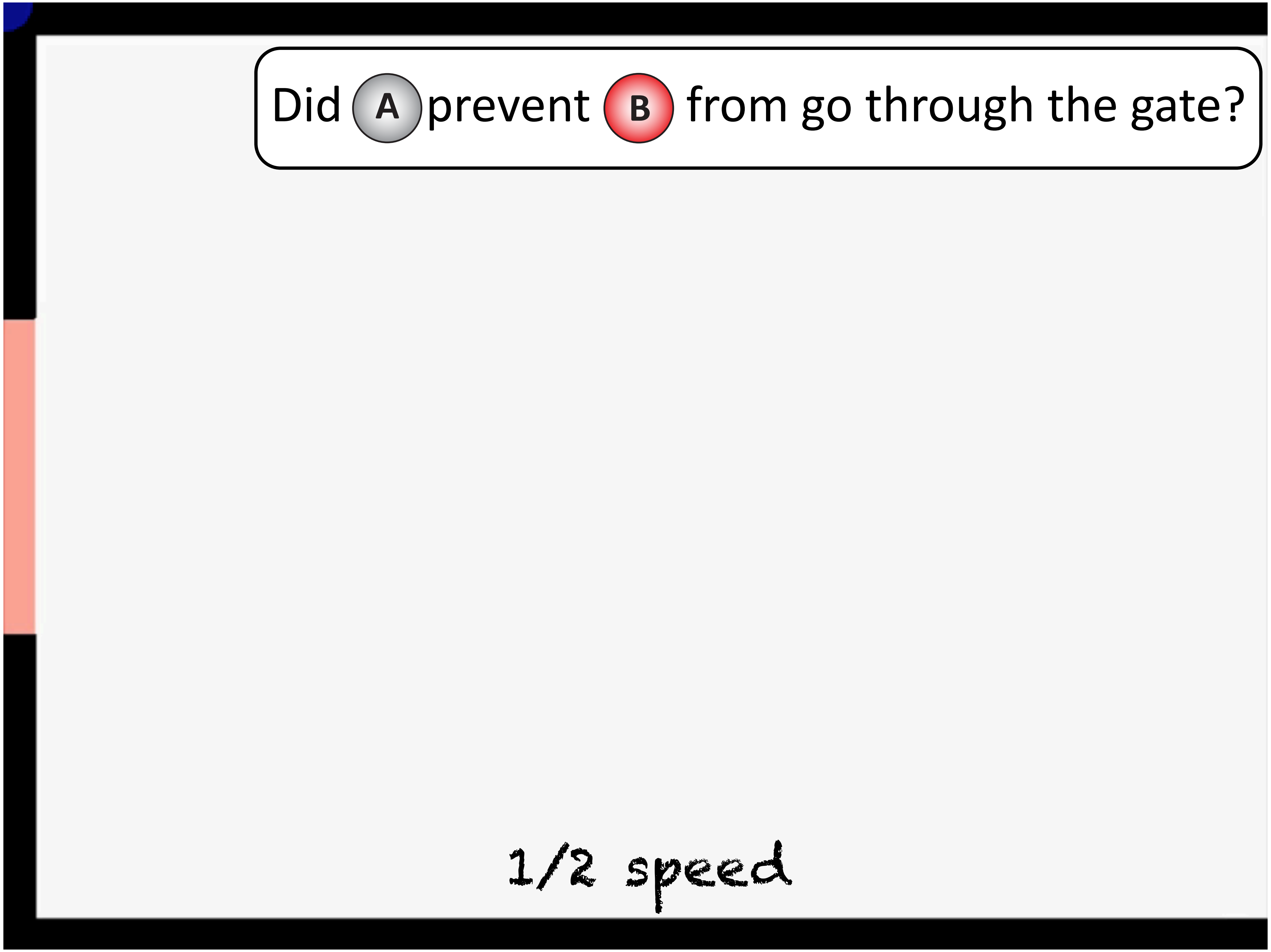
Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum (2017) Eye-tracking causality. *Psychological Science*



Did B completely miss the gate?

1/2 speed

Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum (2017) Eye-tracking causality. *Psychological Science*

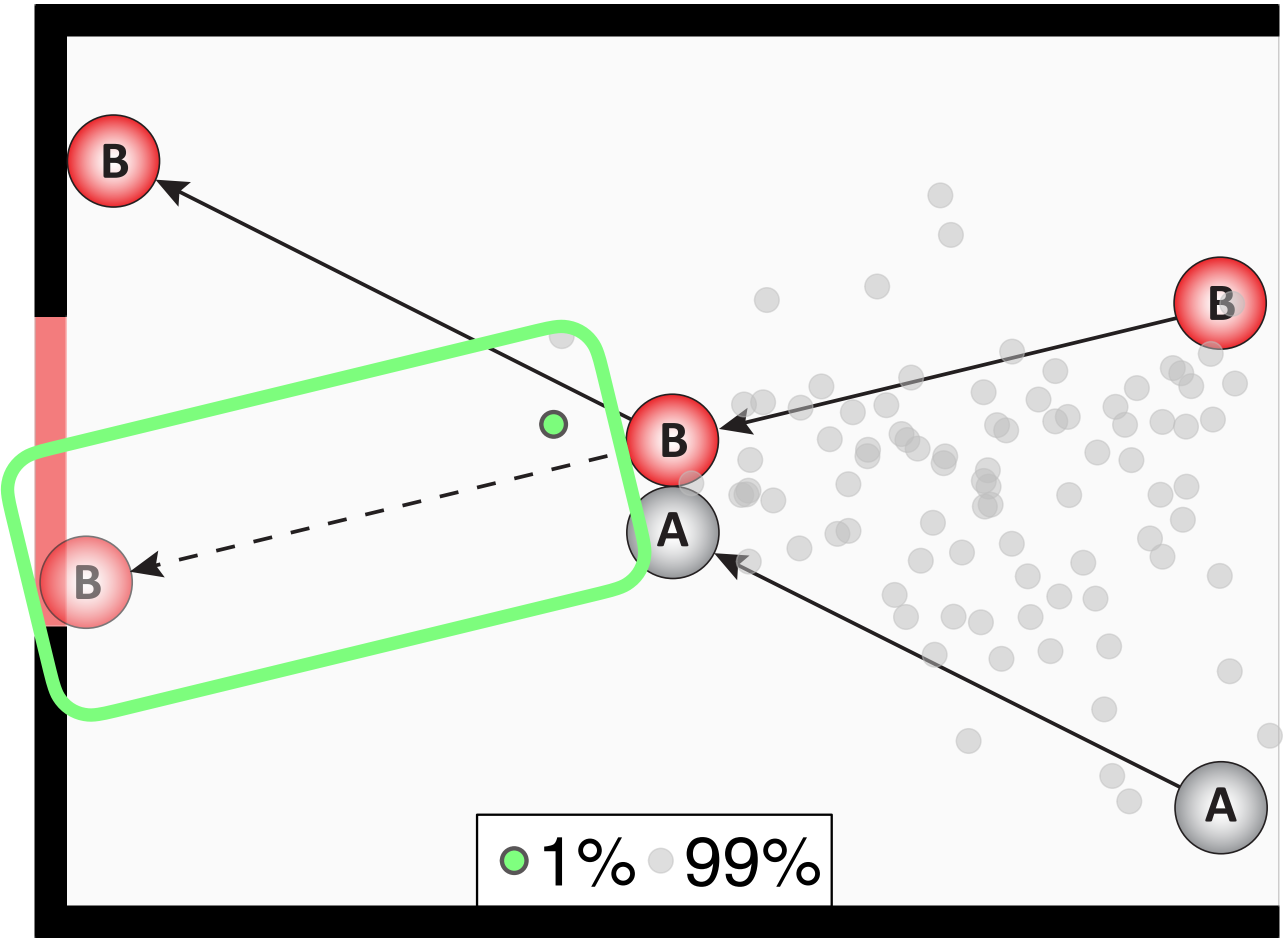


Did A prevent B from go through the gate?

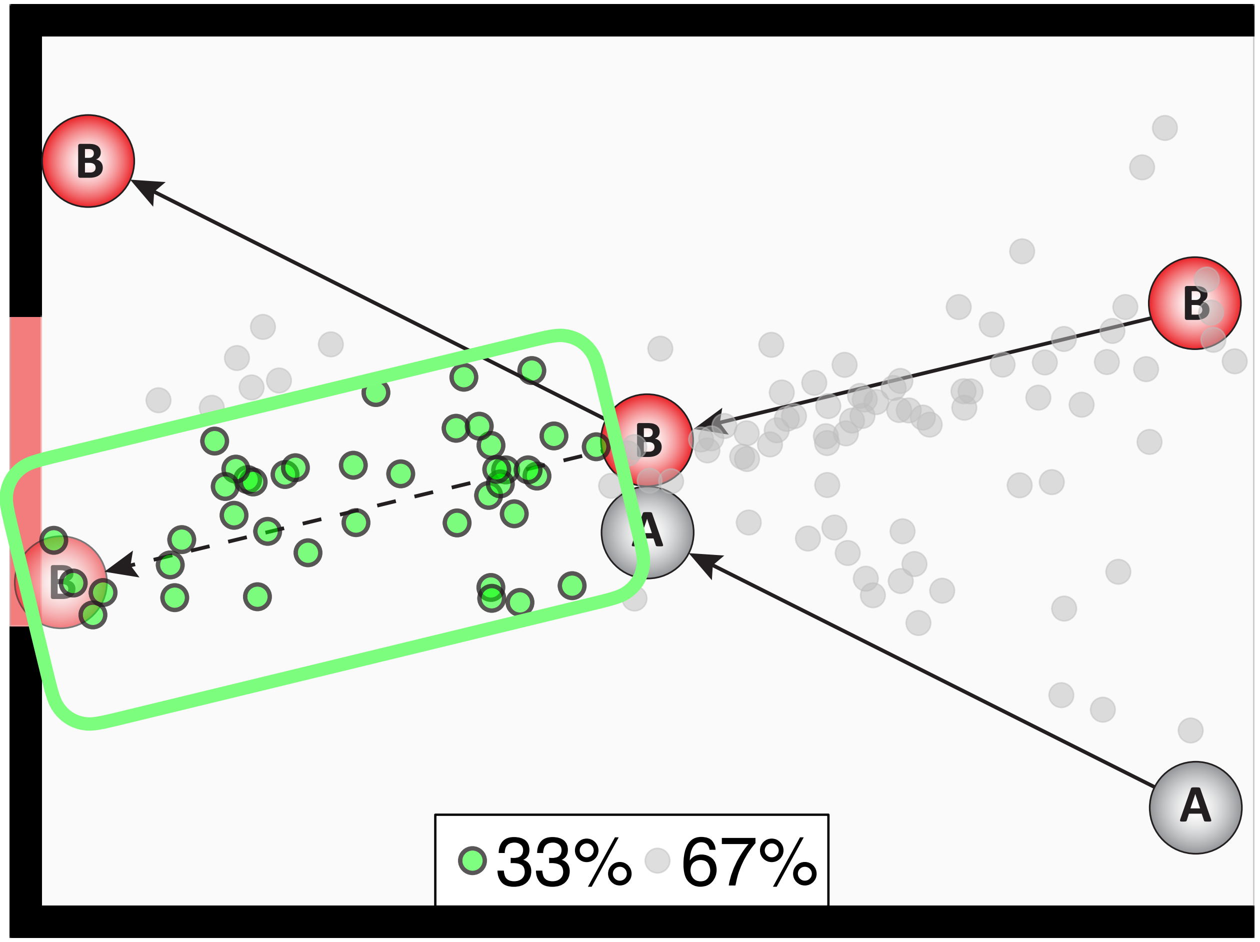
1/2 speed

Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum (2017) Eye-tracking causality. *Psychological Science*

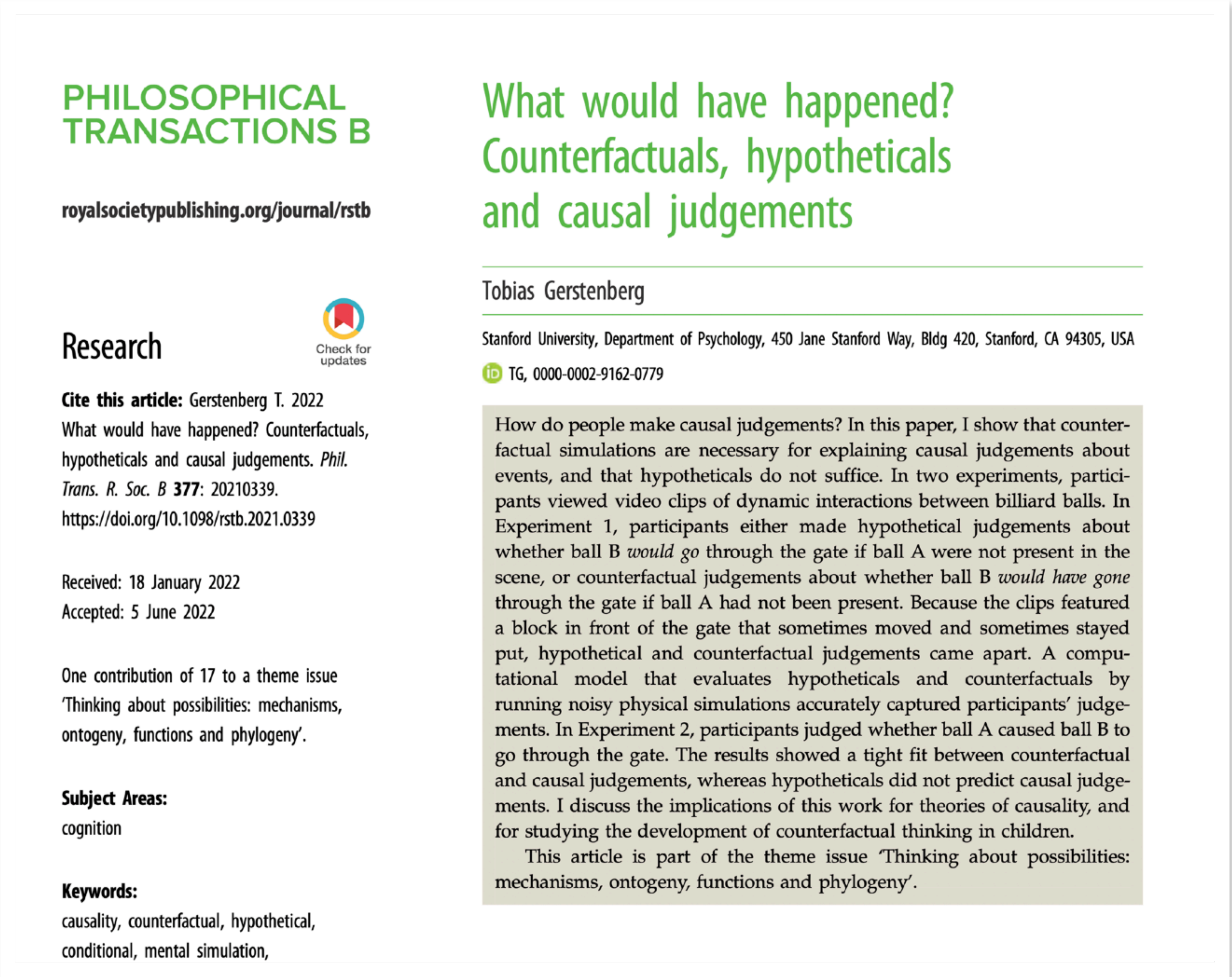
Did B completely miss the gate?



Did A prevent B from go through the gate?



Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum (2017) Eye-tracking causality. *Psychological Science*



PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Check for updates

Cite this article: Gerstenberg T. 2022 What would have happened? Counterfactuals, hypotheticals and causal judgements. *Phil. Trans. R. Soc. B* **377**: 20210339. https://doi.org/10.1098/rstb.2021.0339

Received: 18 January 2022
Accepted: 5 June 2022

One contribution of 17 to a theme issue 'Thinking about possibilities: mechanisms, ontogeny, functions and phylogeny'.

Subject Areas:
cognition

Keywords:
causality, counterfactual, hypothetical, conditional, mental simulation,

What would have happened? Counterfactuals, hypotheticals and causal judgements

Tobias Gerstenberg

Stanford University, Department of Psychology, 450 Jane Stanford Way, Bldg 420, Stanford, CA 94305, USA

TG, 0000-0002-9162-0779

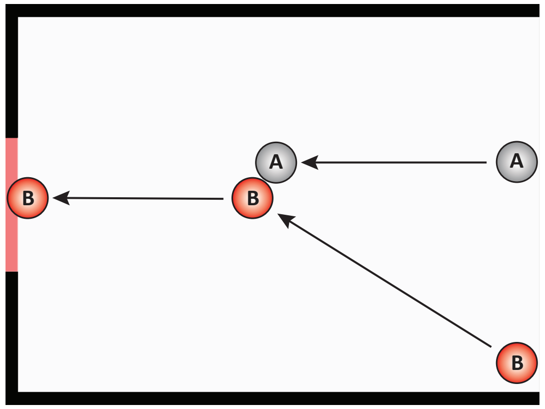
How do people make causal judgements? In this paper, I show that counterfactual simulations are necessary for explaining causal judgements about events, and that hypotheticals do not suffice. In two experiments, participants viewed video clips of dynamic interactions between billiard balls. In Experiment 1, participants either made hypothetical judgements about whether ball B *would go* through the gate if ball A were not present in the scene, or counterfactual judgements about whether ball B *would have gone* through the gate if ball A had not been present. Because the clips featured a block in front of the gate that sometimes moved and sometimes stayed put, hypothetical and counterfactual judgements came apart. A computational model that evaluates hypotheticals and counterfactuals by running noisy physical simulations accurately captured participants' judgements. In Experiment 2, participants judged whether ball A caused ball B to go through the gate. The results showed a tight fit between counterfactual and causal judgements, whereas hypotheticals did not predict causal judgements. I discuss the implications of this work for theories of causality, and for studying the development of counterfactual thinking in children.

This article is part of the theme issue 'Thinking about possibilities: mechanisms, ontogeny, functions and phylogeny'.

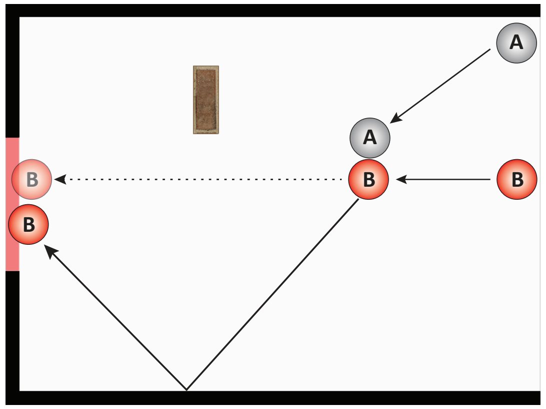
Do you really **need counterfactuals** to explain causal judgments?

Gerstenberg, T. (2022). What would have happened? Counterfactuals, hypotheticals, and causal judgments. *Philosophical Transactions of the Royal Society B: Biological Sciences.*

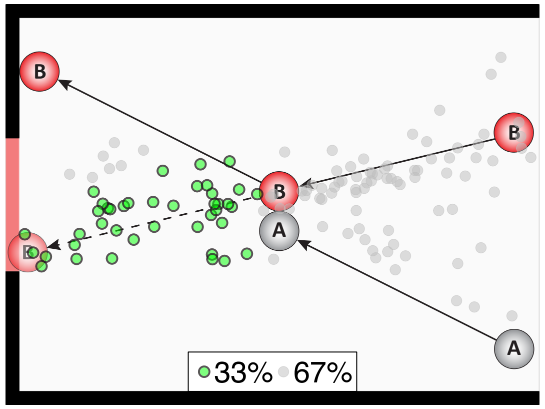
Counterfactual simulation model of causal judgment



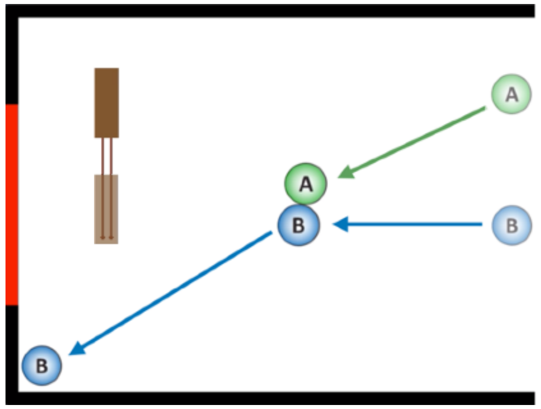
- causal judgments are well-explained by the observer's beliefs about **whether** the candidate cause made a difference to the outcome



- counterfactual contrasts are **necessary** for explaining people's causal judgments



- people **spontaneously** engage in counterfactual simulation when making causal judgments



- **counterfactuals** (not hypotheticals) explain causal judgments

Counterfactual simulation model of causal judgment

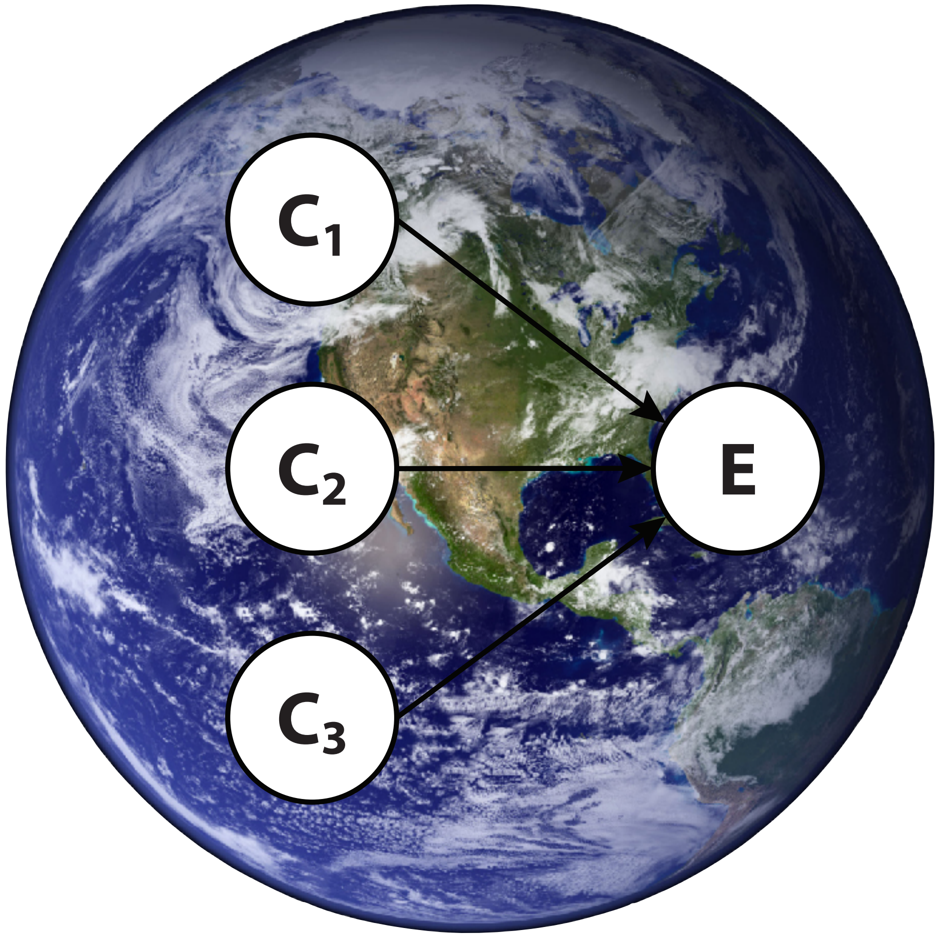
Did **E** go into the gate because of **B**?

event
causality

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021) A counterfactual simulation model of causal judgments for physical events. *Psychological Review*

A computational framework for understanding responsibility

What causal role did the action play?



Intuitive theory of how **the world** works

What does the action reveal about the person?

Intuitive theory of how **people** work

Gerstenberg & Tenenbaum (2017) Intuitive Theories. *Oxford Handbook of Causal Reasoning*

Goodman, Tenenbaum, & Gerstenberg (2015) Concepts in a probabilistic language of thought. *The Conceptual Mind: New Directions in the Study of Concepts*

A computational framework for understanding responsibility

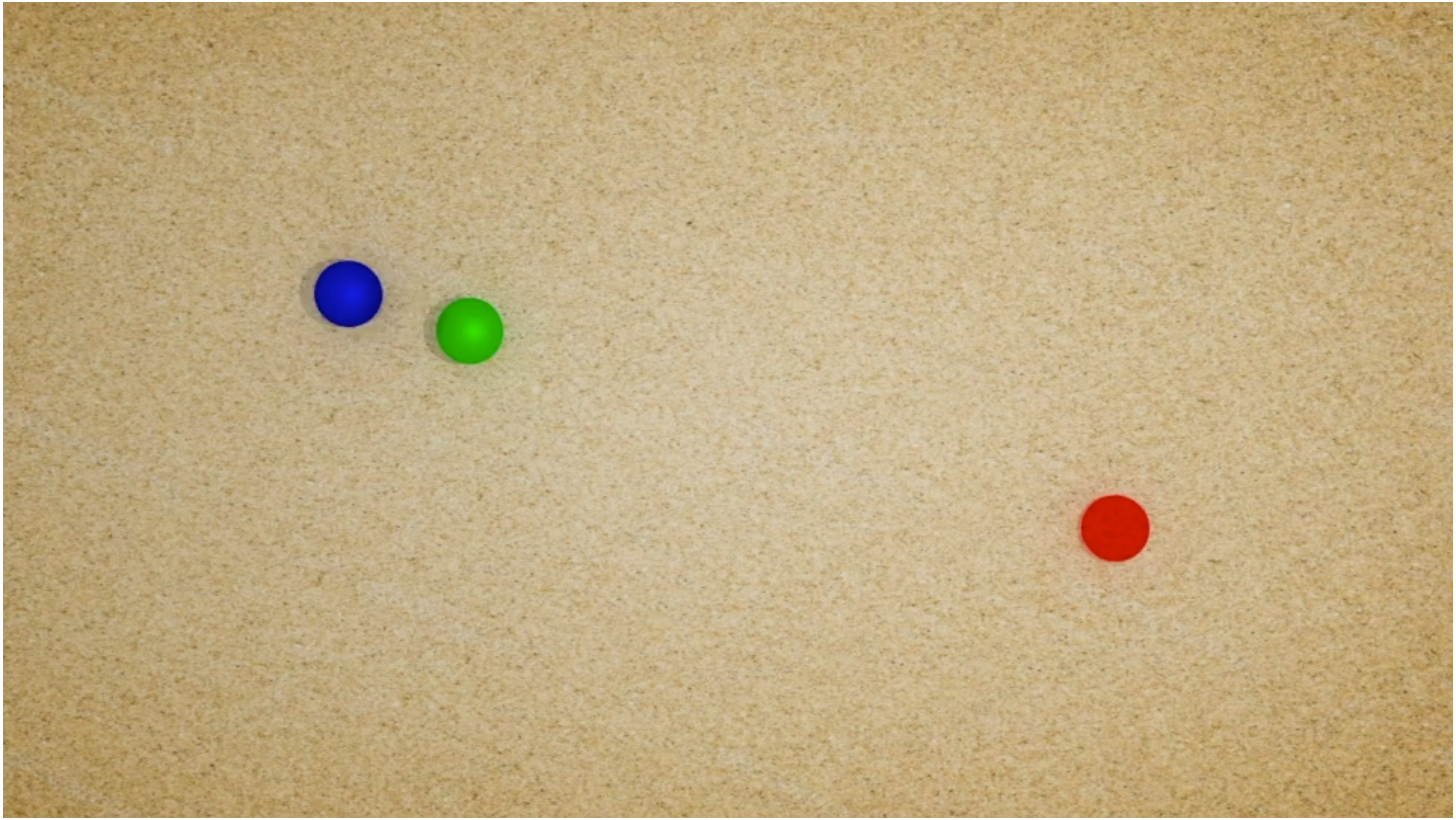
What does the action
reveal about the person?

Intuitive theory of
how **people** work

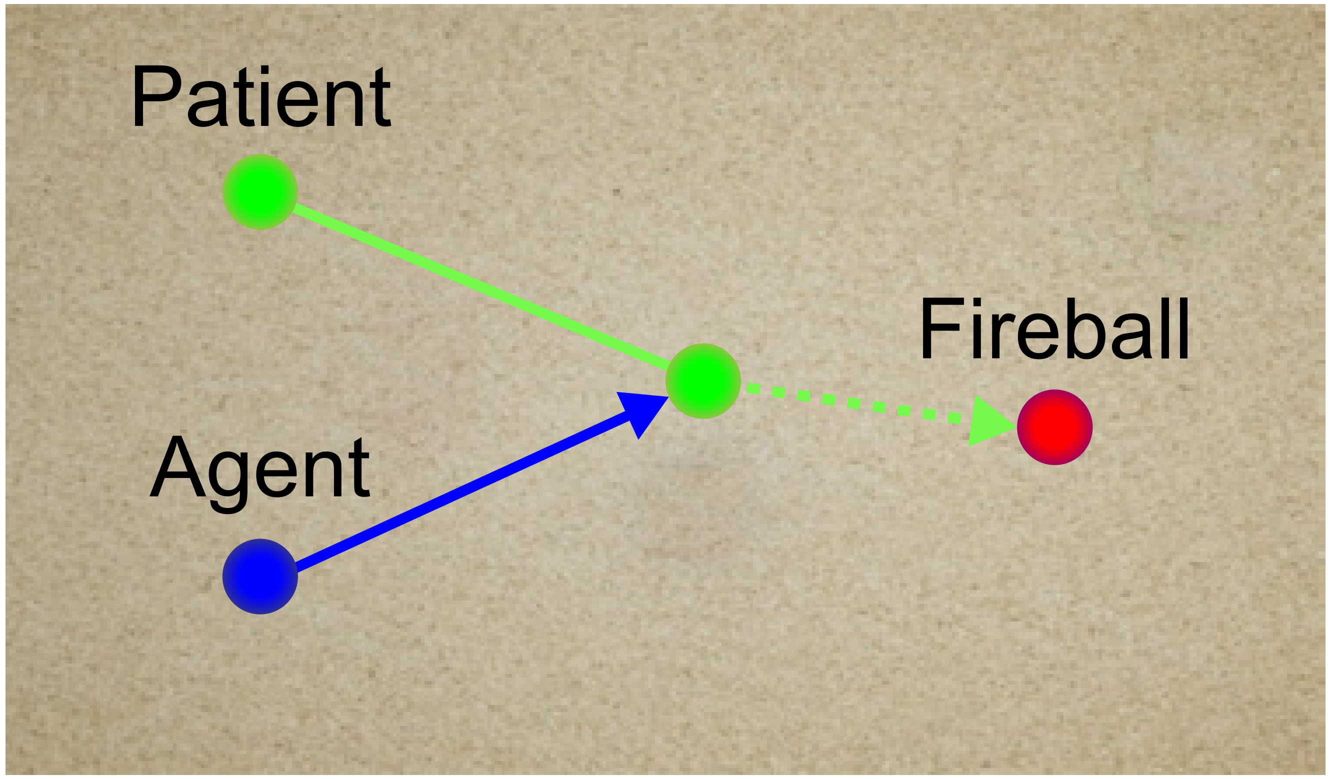
Gerstenberg & Tenenbaum (2017) Intuitive Theories. *Oxford Handbook of Causal Reasoning*

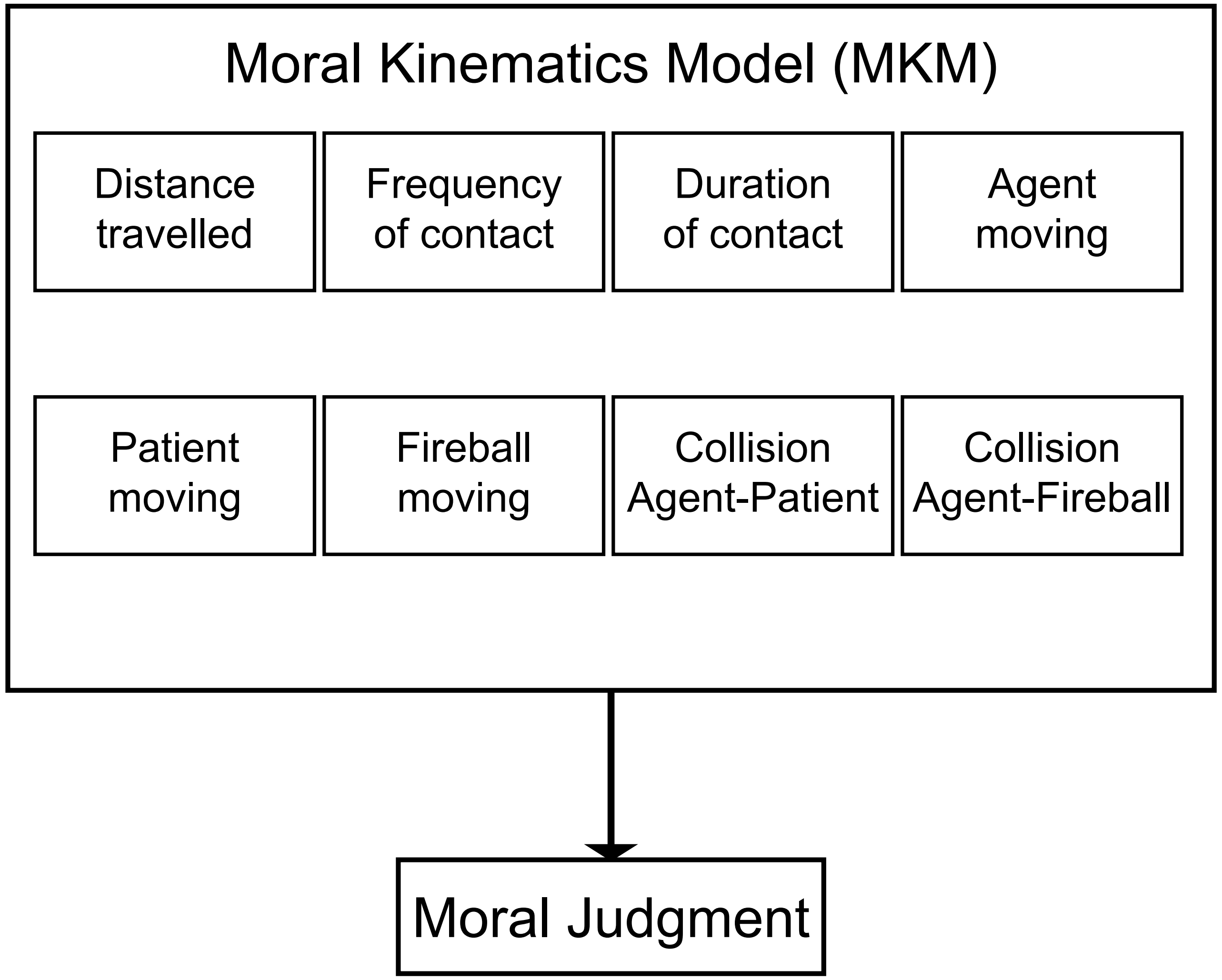
Goodman, Tenenbaum, & Gerstenberg (2015) Concepts in a probabilistic language of thought. *The Conceptual Mind: New Directions in the Study of Concepts*

To what extent was **Blue** responsible that **Green** got harmed?



Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg (2021) Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*





Moral Kinematics Model (MKM)

Distance travelled

Frequency of contact

Duration of contact

Agent moving

Patient moving

Fireball moving

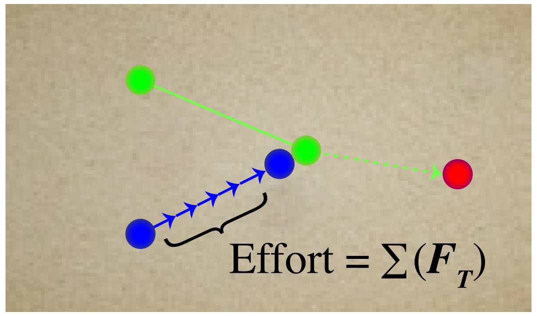
Collision Agent-Patient

Collision Agent-Fireball

Moral Judgment

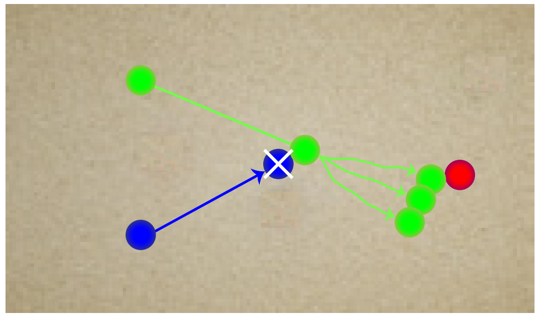
Iliev, Sachdeva, & Medin (2012) Moral kinematics: The role of physical factors in moral judgments. *Memory & Cognition*

Intuitive Psychology

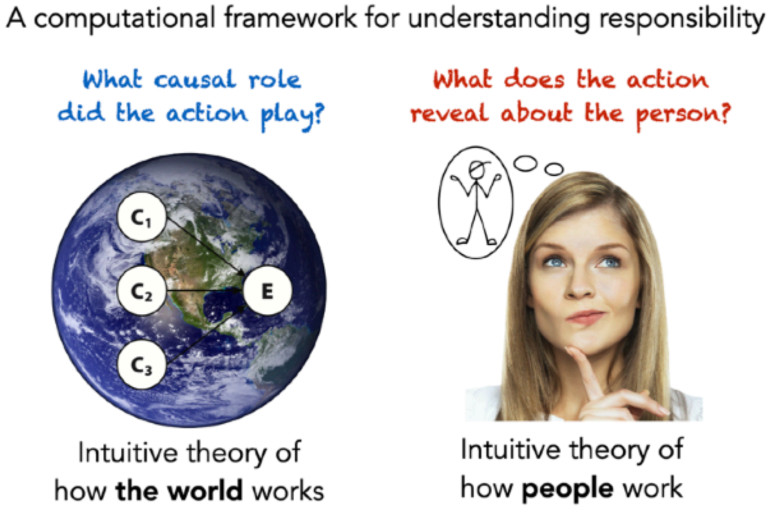


Action cost → Desire to harm

Intuitive Physics



Counterfactual simulation → Cause of harm

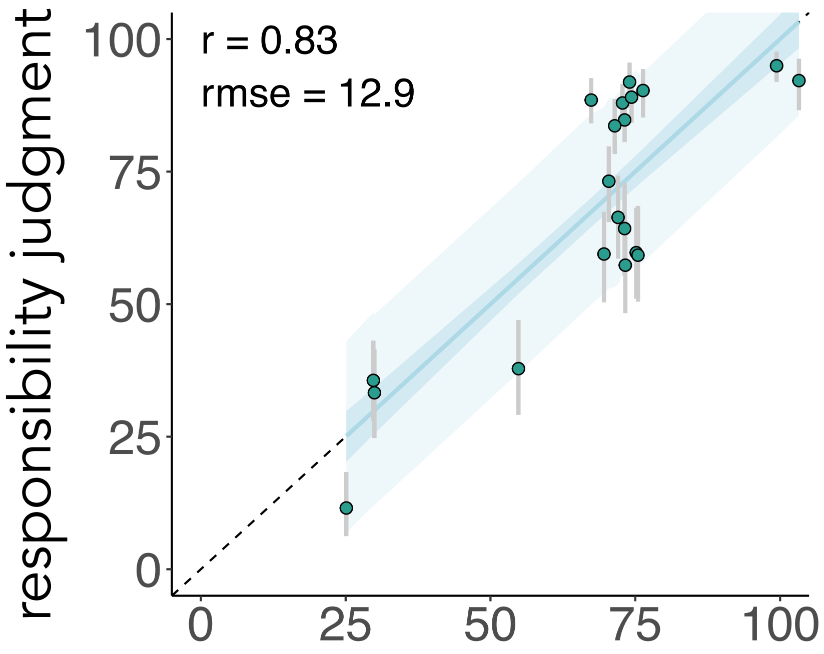

A computational framework for understanding responsibility

What causal role did the action play?

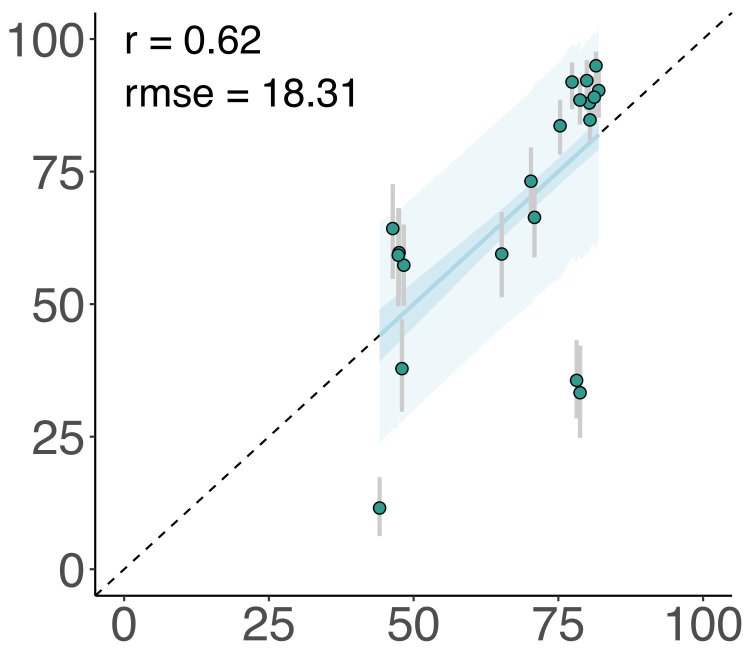
What does the action reveal about the person?

Intuitive theory of how **the world** works

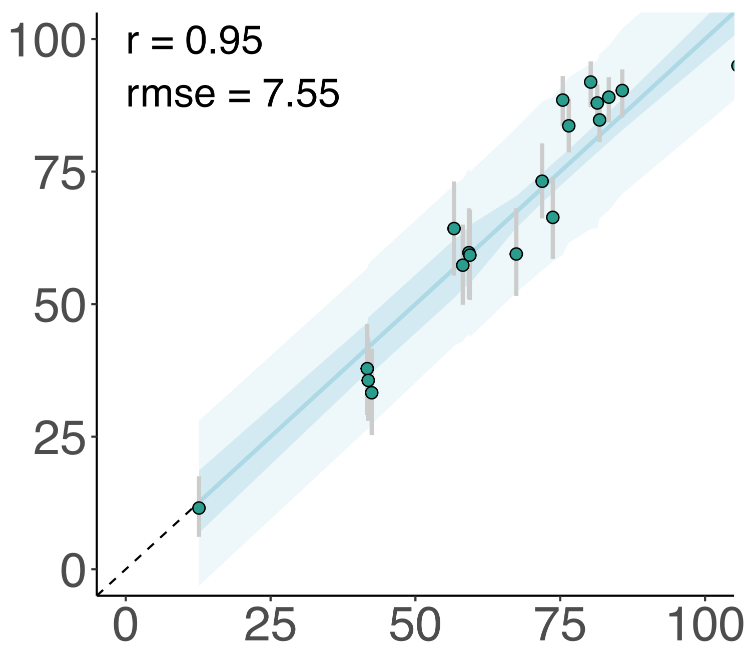
Intuitive theory of how **people** work



effort



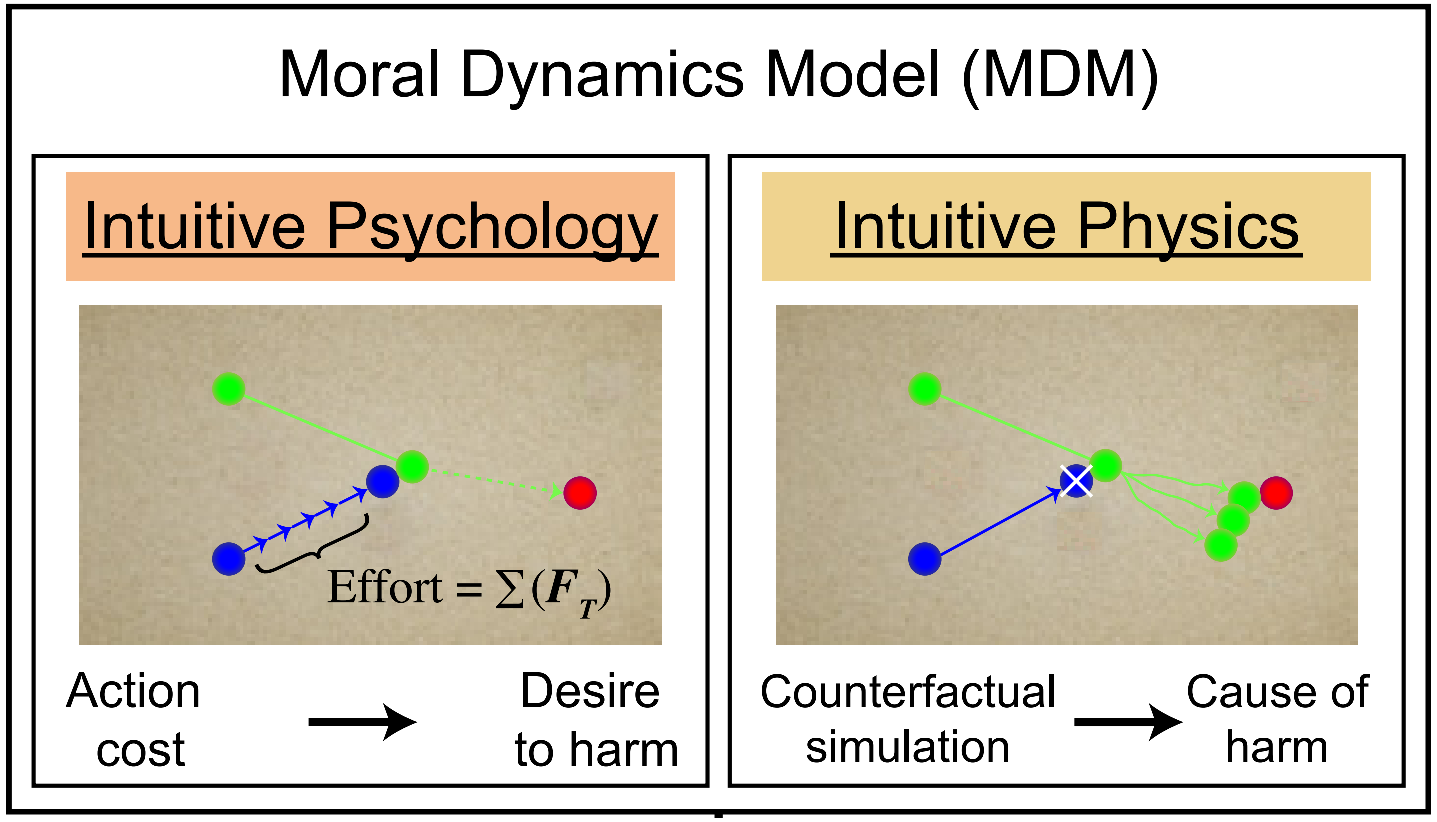
causality



effort + causality

Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg (2021) Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*

But ...



Moral Dynamics Model (MDM)

Intuitive Psychology

Effort = $\sum(F_T)$

Action cost → Desire to harm

Intuitive Physics

Counterfactual simulation → Cause of harm

- no real model of agents
- no model of intention inference
- counterfactual simulation is purely physical

Sarah Wu

Shruti Sridhar

Experiment 1

planning actions

Experiment 2

helping / hindering

Sarah Wu

Shruti Sridhar

Experiment 1

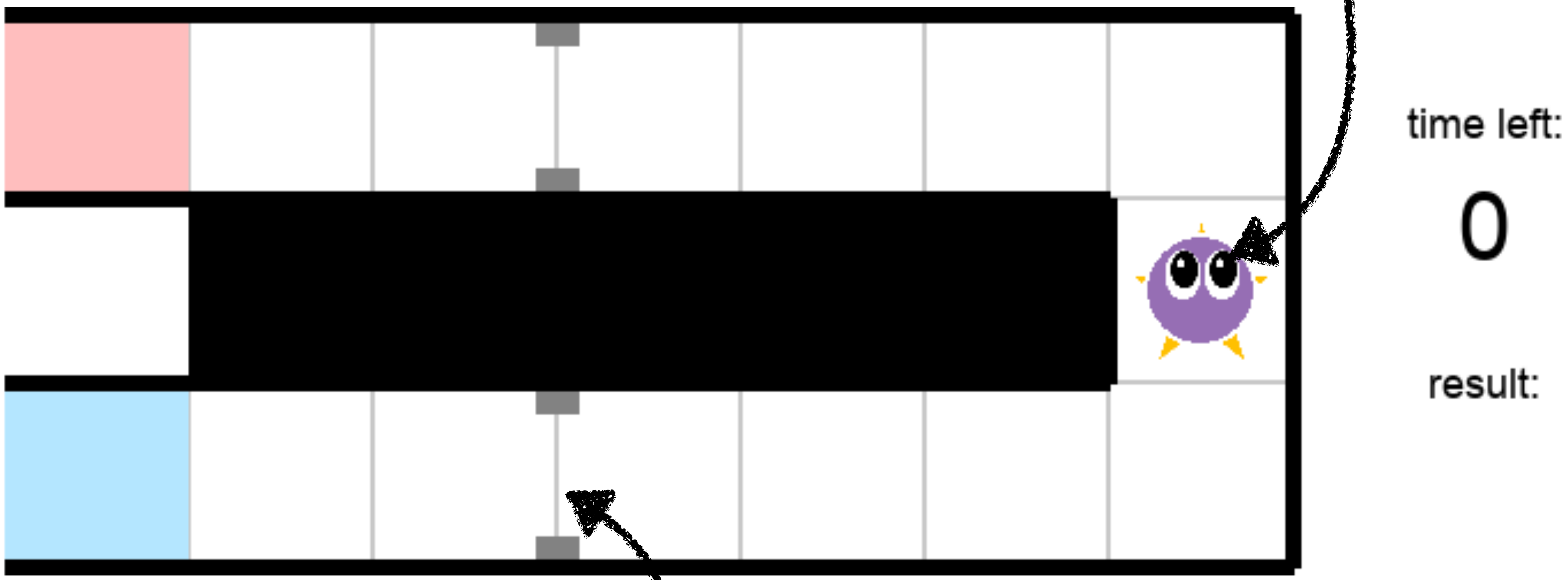
planning actions

Experiment 2

helping / hindering

the agent needs to decide which path to take

the agent wins if it reaches the star in time

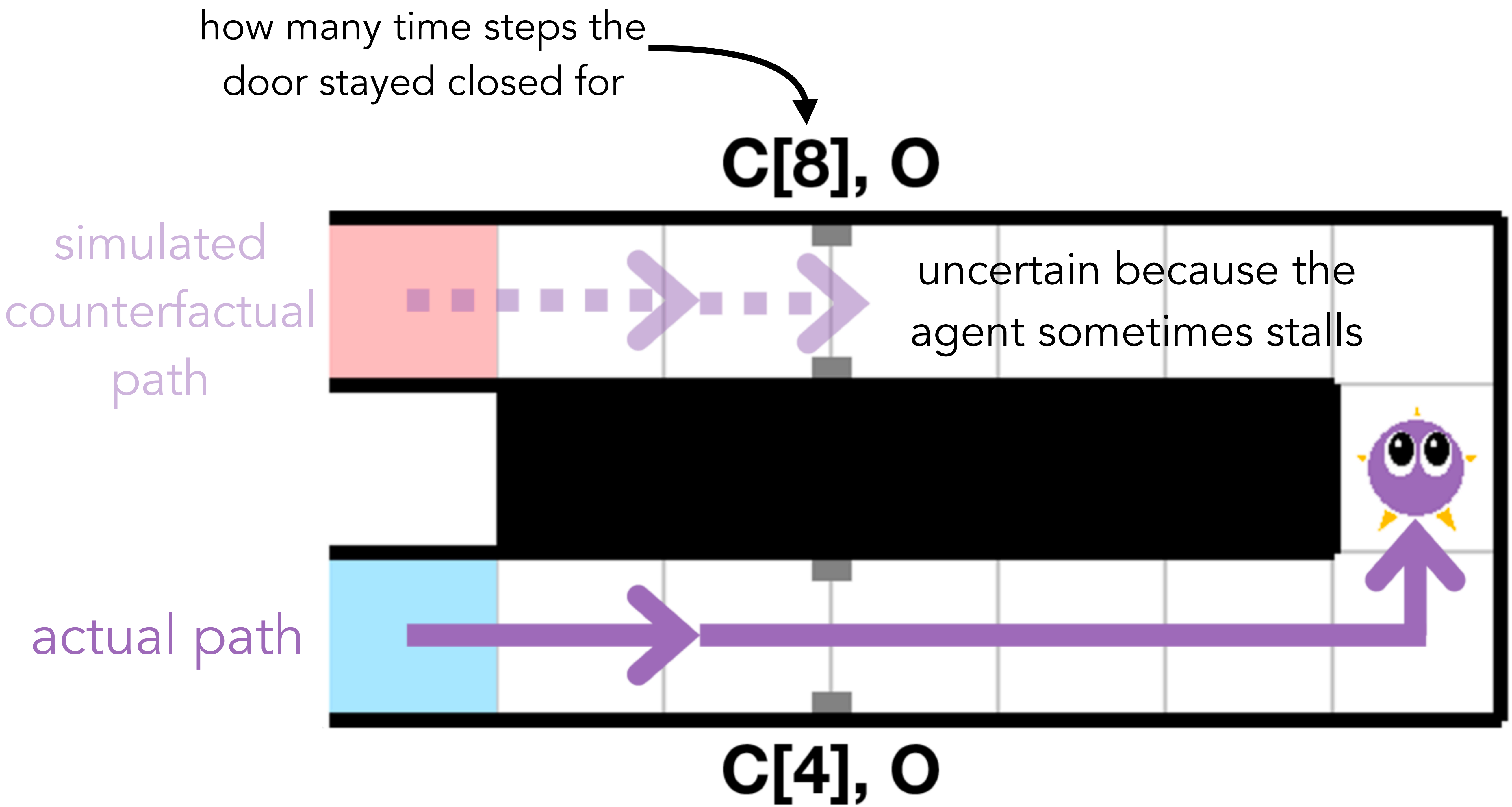


doors can randomly open or close

Did the agent win because it took the blue path this time?

Wu, Sridhar, & Gerstenberg (2022) That was close! A counterfactual simulation model of causal judgments about decisions. *CogSci Proceedings*

Counterfactual simulation model of causal judgment



how many time steps the door stayed closed for

C[8], O

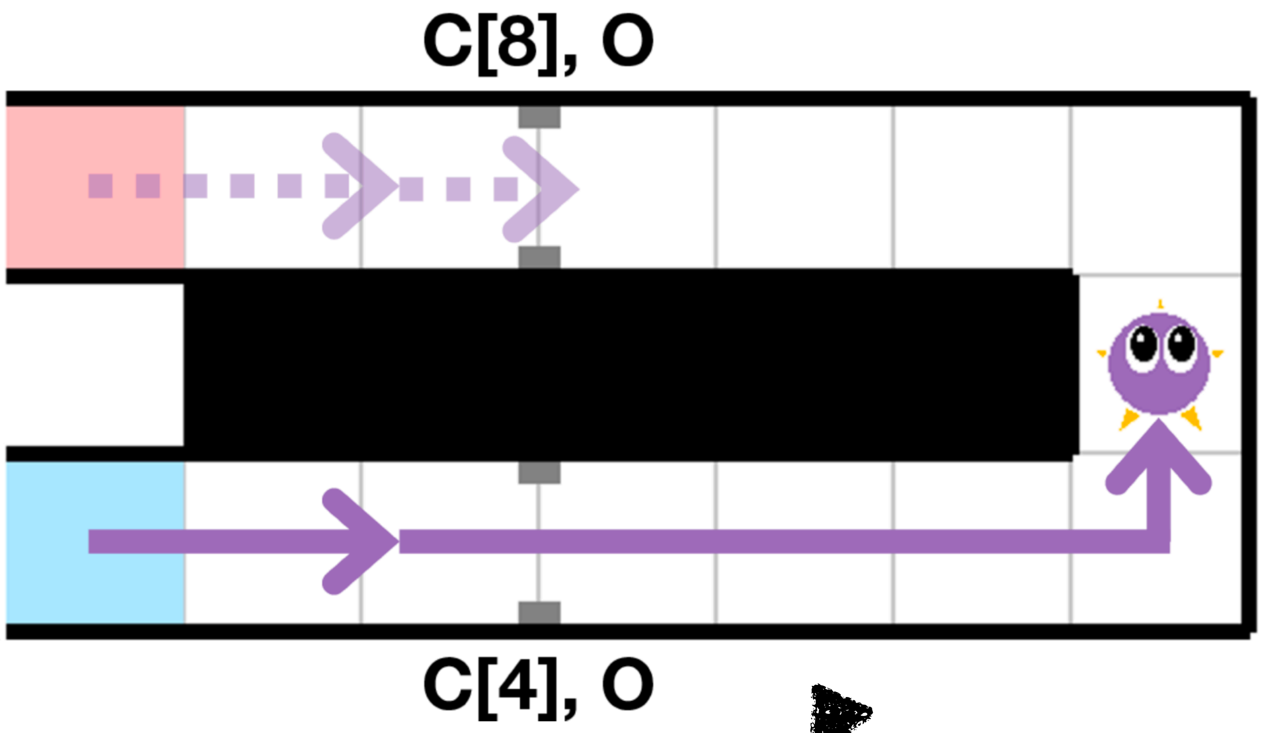
simulated counterfactual path

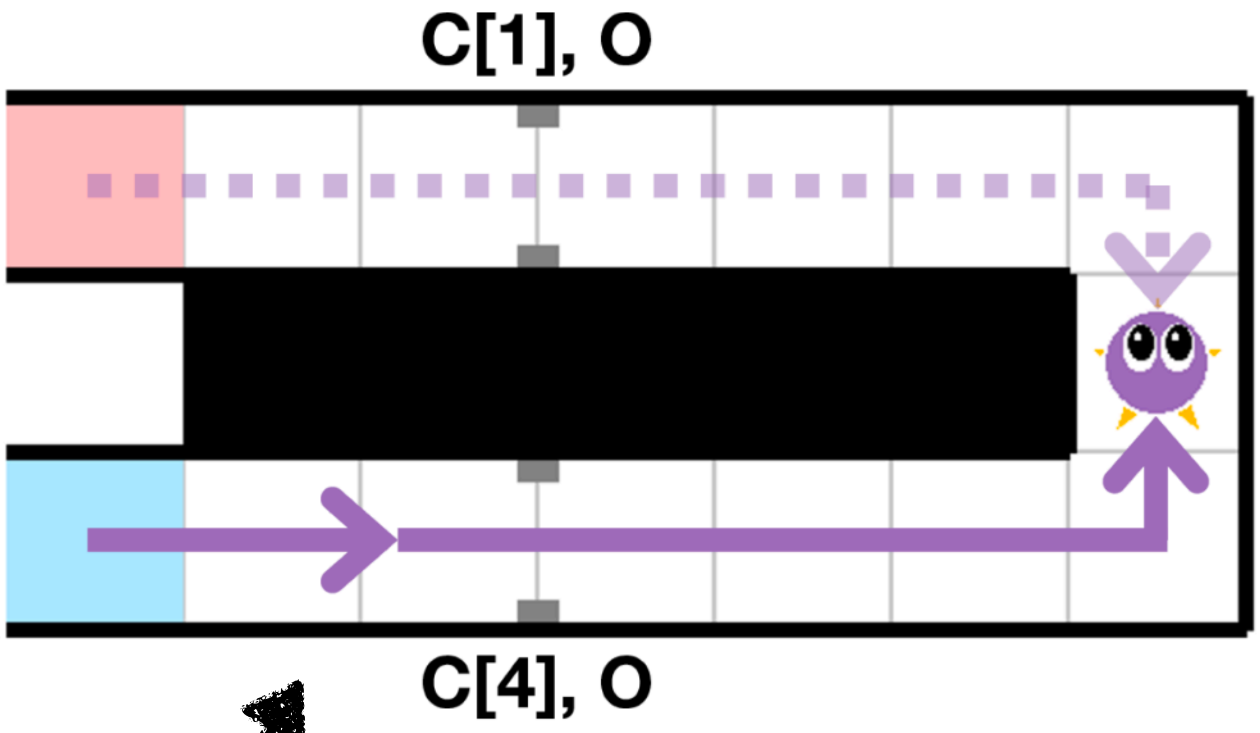
uncertain because the agent sometimes stalls

actual path

C[4], O

Counterfactual simulation model of causal judgment





yes

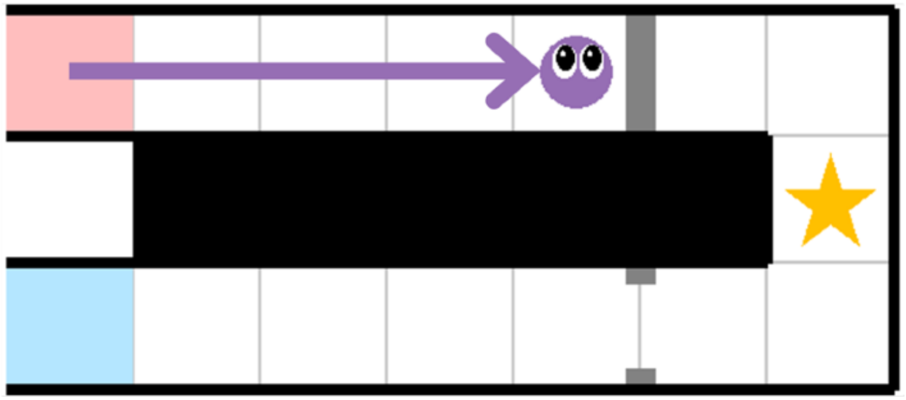
no

Did the agent win because it took the blue path this time?

Counterfactual simulation model of causal judgment

actual situation:

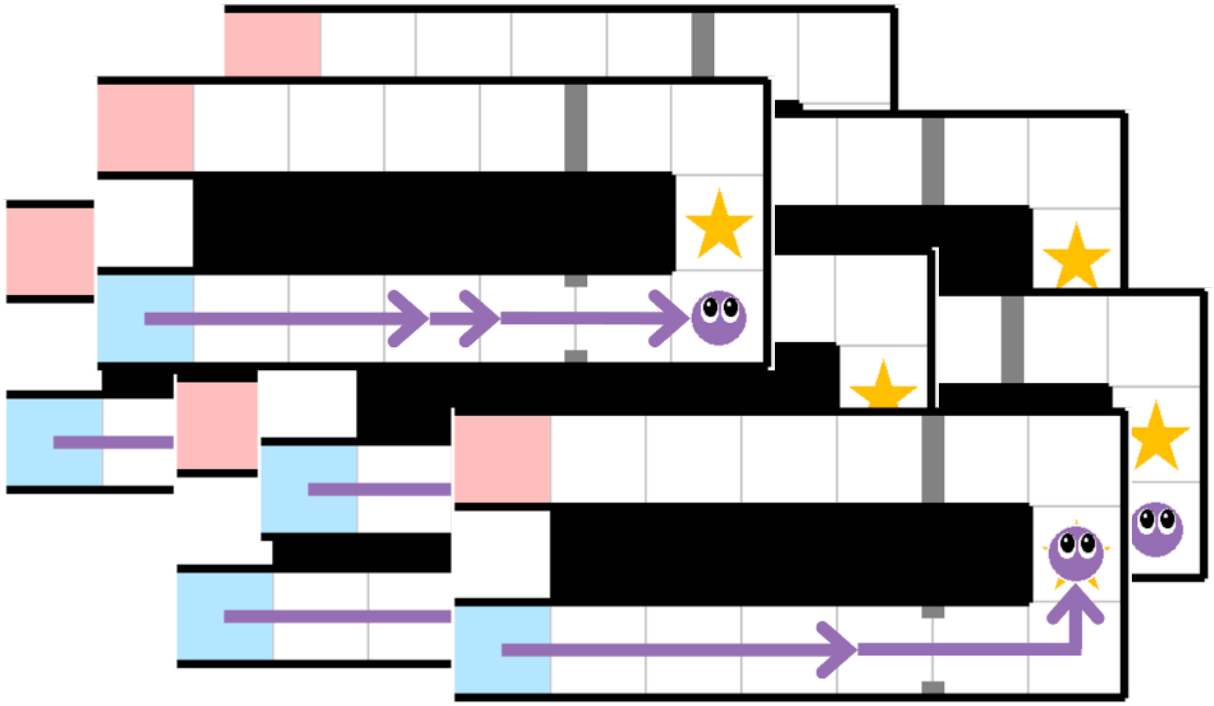
red path, loss



counterfactual simulations:

what would have happened

if the agent had taken the blue path



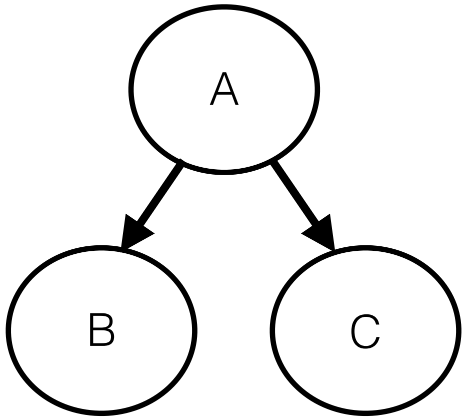
counterfactual outcome: 68% success

Causal judgments as counterfactual contrasts over generative models

Generative model

causal Bayes net

structural equations



$$B = A$$

$$C = A$$

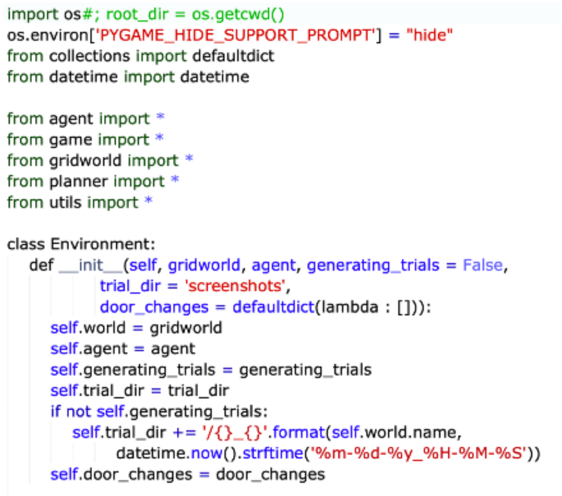
Counterfactual intervention

do() operator

Pearl, J. (2000). *Causality: Models, reasoning and inference*

Generative model

probabilistic program



```
import os#; root_dir = os.getcwd()
os.environ['PYGAME_HIDE_SUPPORT_PROMPT'] = "hide"
from collections import defaultdict
from datetime import datetime

from agent import *
from game import *
from gridworld import *
from planner import *
from utils import *

class Environment:
    def __init__(self, gridworld, agent, generating_trials = False,
                 trial_dir = 'screenshots',
                 door_changes = defaultdict(lambda : [])):
        self.world = gridworld
        self.agent = agent
        self.generating_trials = generating_trials
        self.trial_dir = trial_dir
        if not self.generating_trials:
            self.trial_dir += '/{}_{}'.format(self.world.name,
                    datetime.now().strftime('%m-%d-%y_%H-%M-%S'))
        self.door_changes = door_changes
```

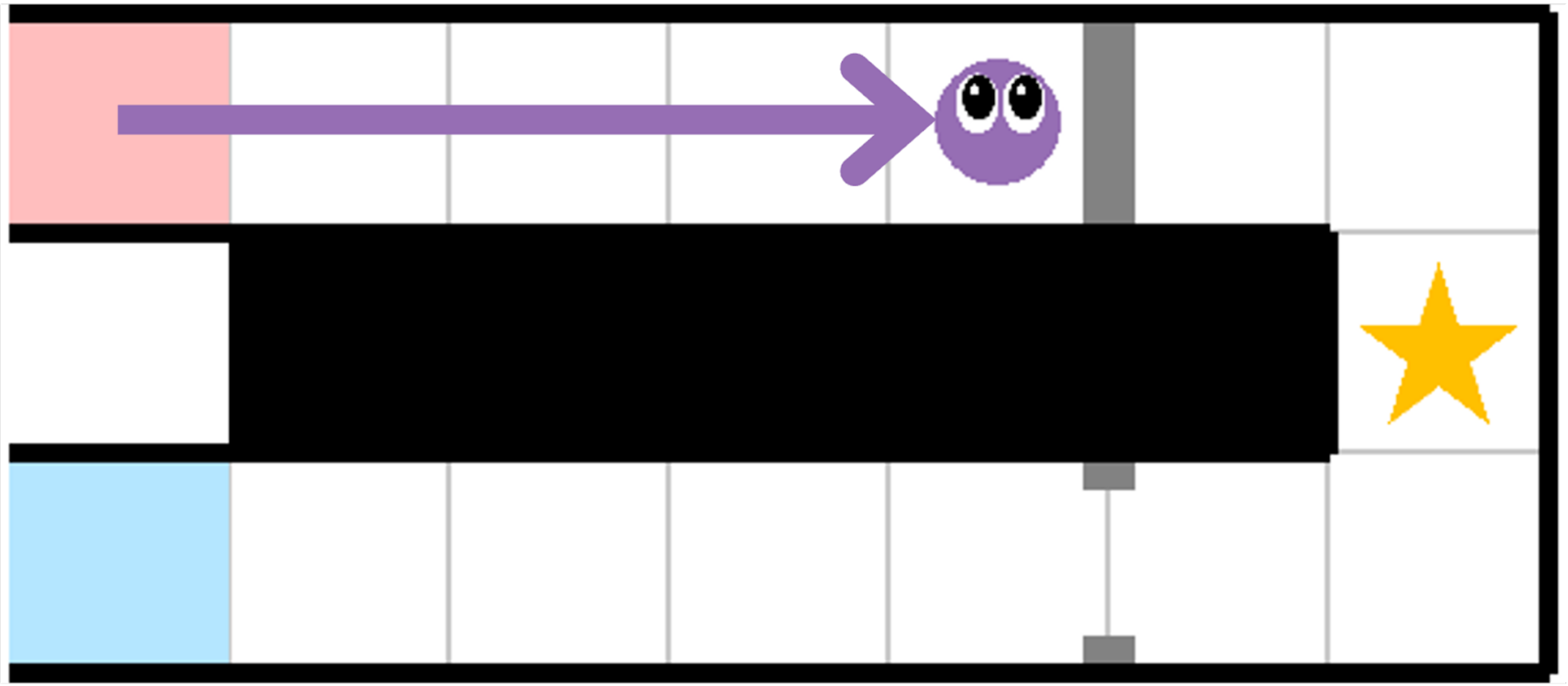
Counterfactual intervention

change(agent) operator

Chater & Oaksford (2013) Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*

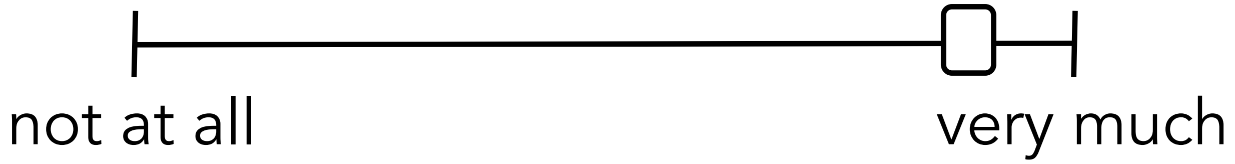
Goodman, Tenenbaum, & Gerstenberg (2015) Concepts in a probabilistic language of thought. *The Conceptual Mind: New Directions in the Study of Concepts*

Cause

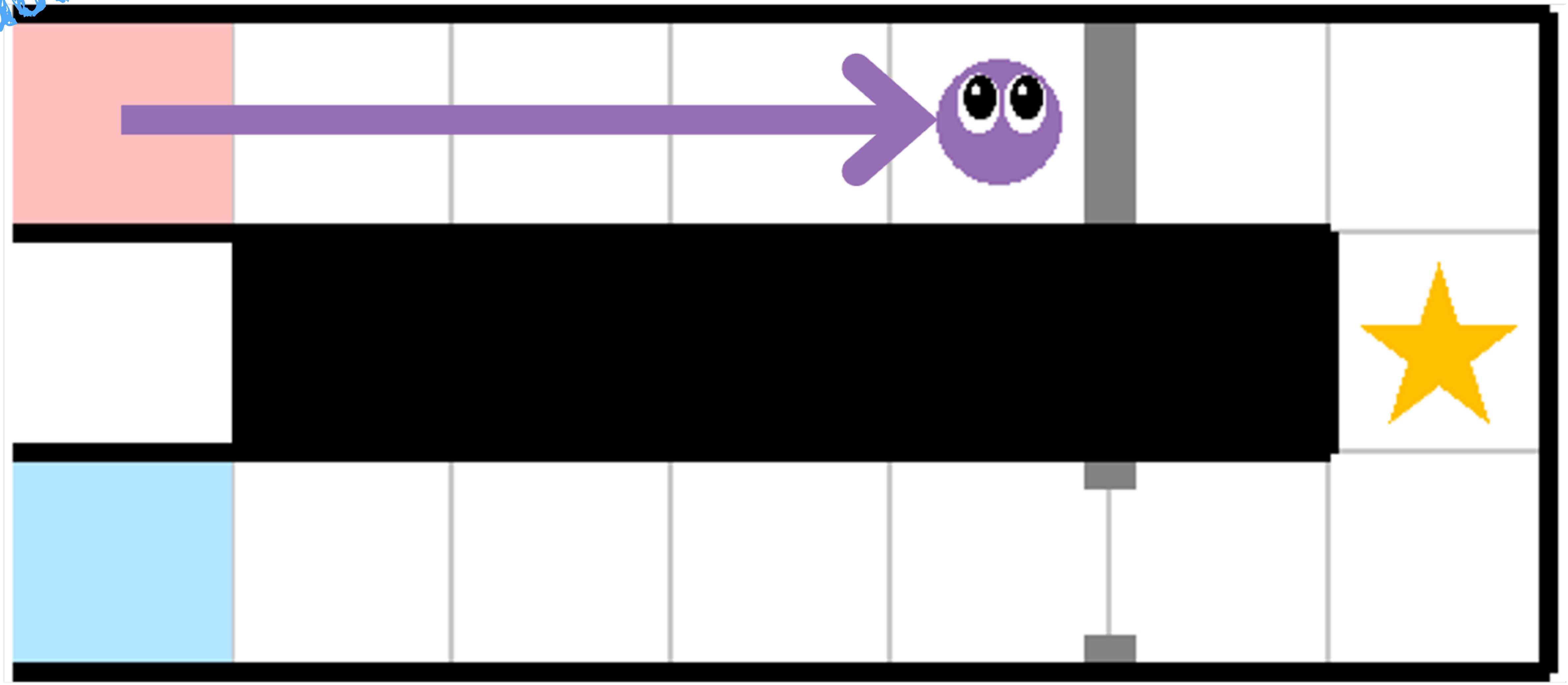


To what extent do you agree with the following statement?

"The player lost because they took the red path this time."

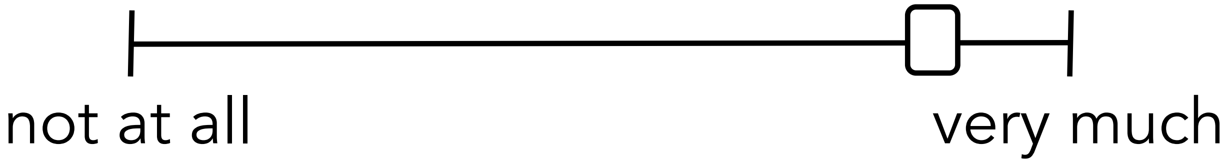


counter factual

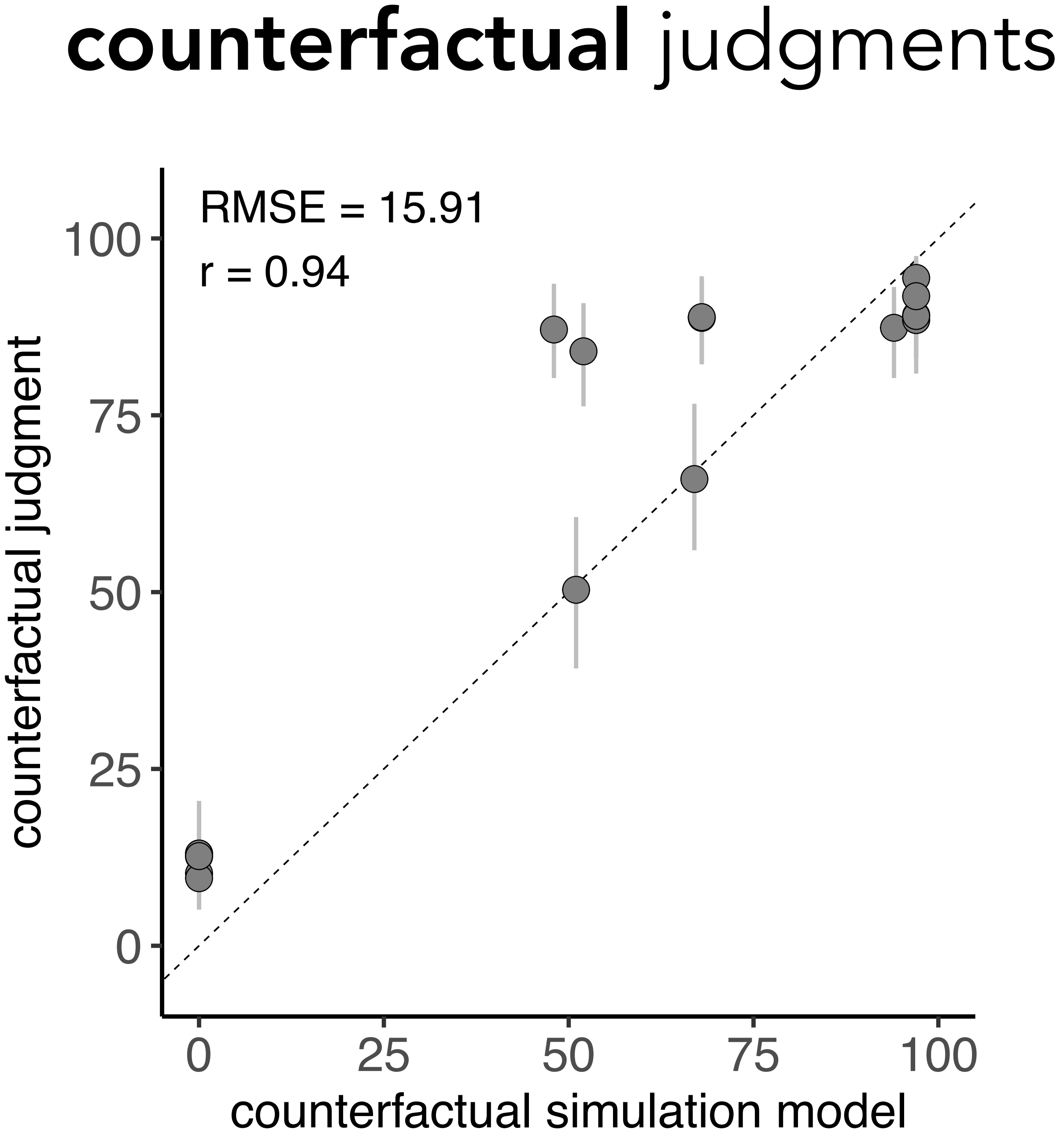


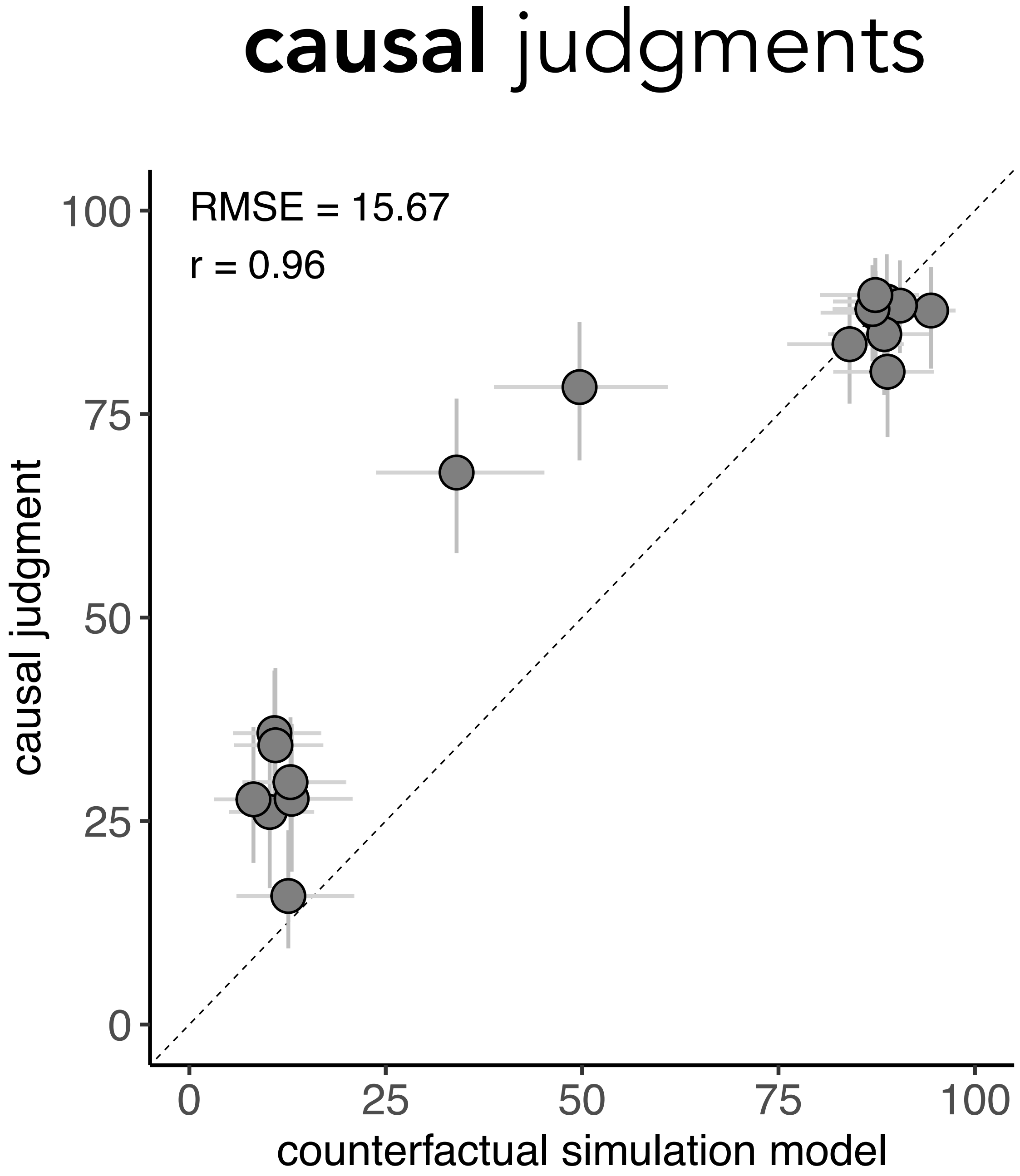
To what extent do you agree with the following statement?

"If the player had taken the blue path this time, the would have won."



(n = 50 each)





CSM captures counterfactual judgments

counterfactuals explain causal judgments

Sarah Wu

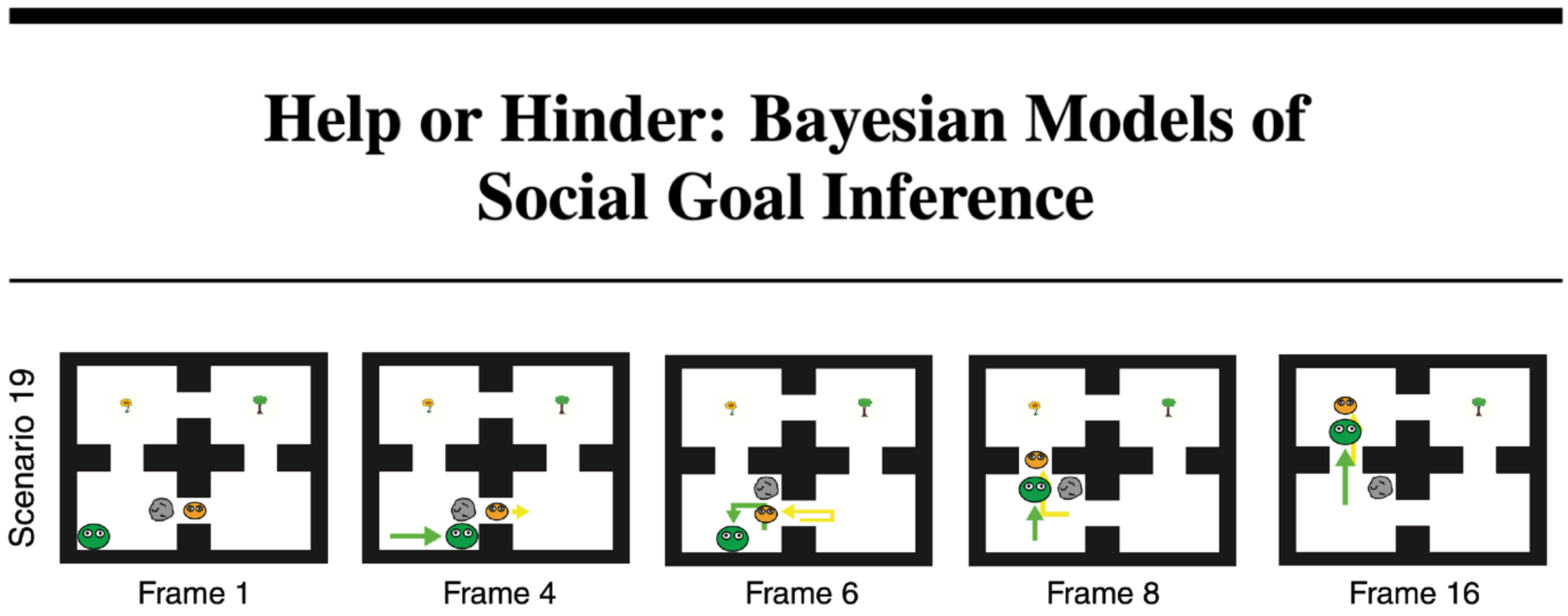
Shruti Sridhar

Experiment 1

planning actions

Experiment 2

helping / hindering



Help or Hinder: Bayesian Models of Social Goal Inference

child "helping"
with the groceries

$+U(\bullet)$ $-U(\bullet)$

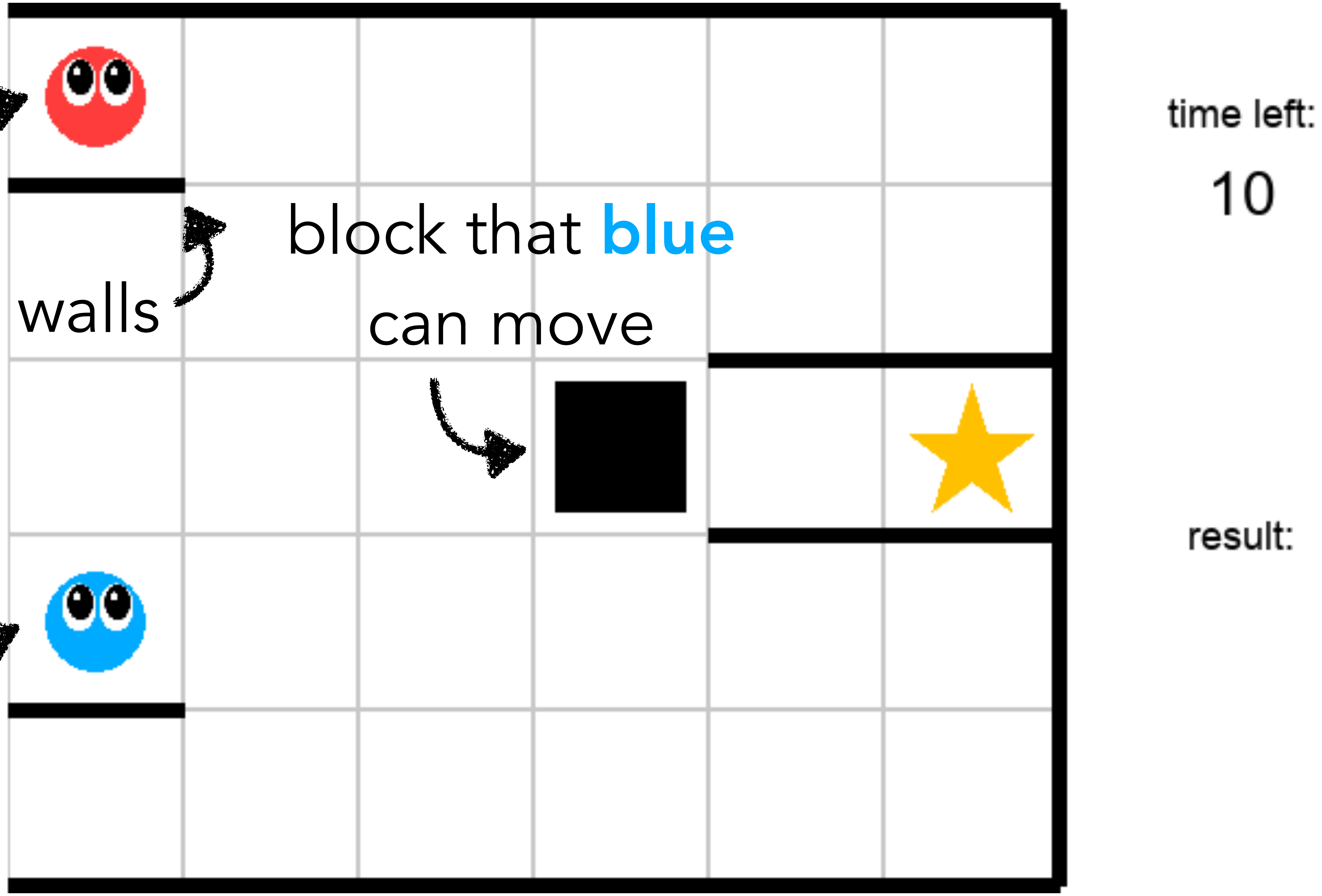
intending to help

intending to hinder

counterfactuals
needed!

intending to help/hinder **vs.** actually helping/hindering

Ullman, Tenenbaum, Baker, Macindoe, Evans, & Goodman (2009) Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems* wants to get to the star



time left:

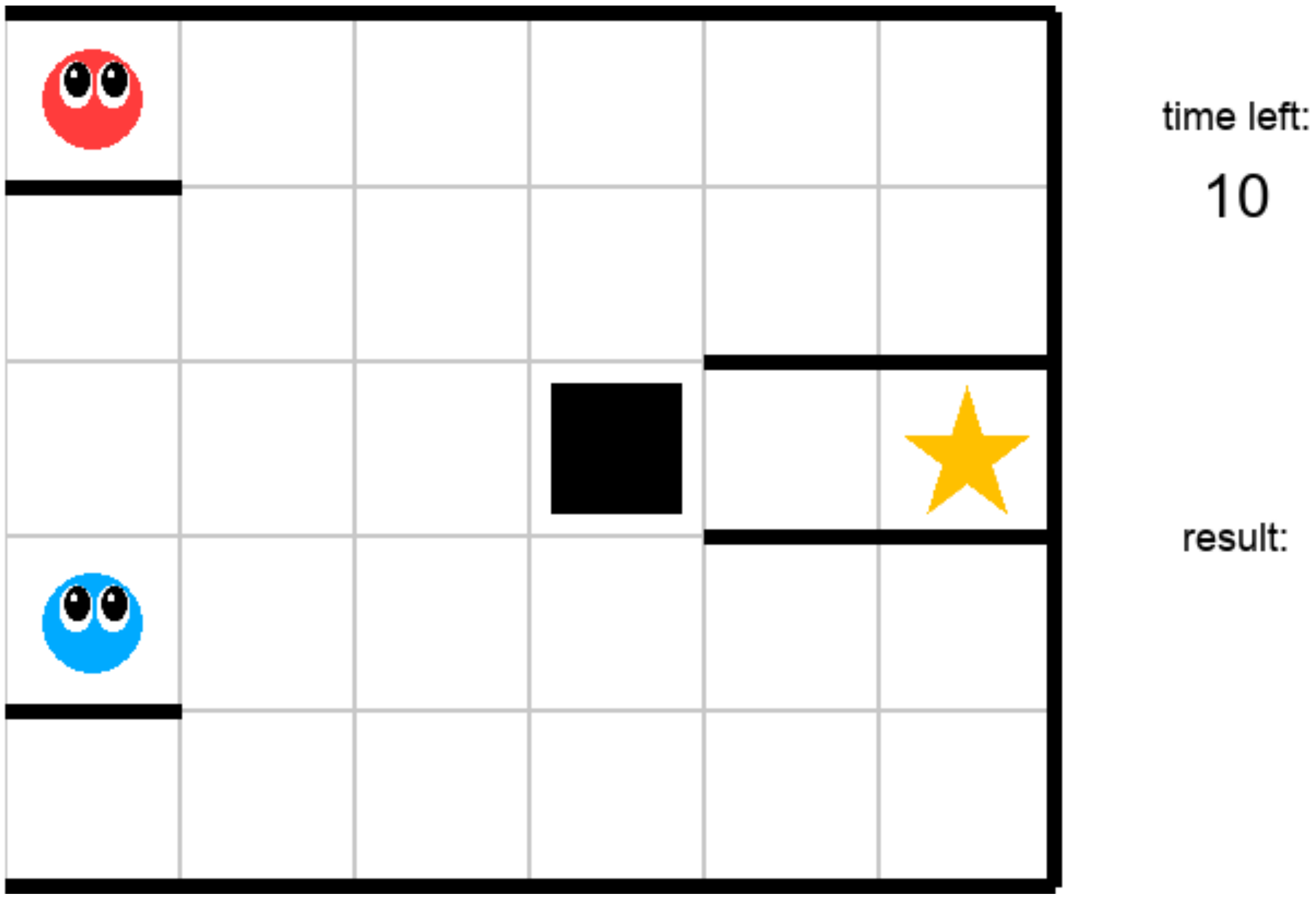
10

result:

static walls

block that **blue** can move

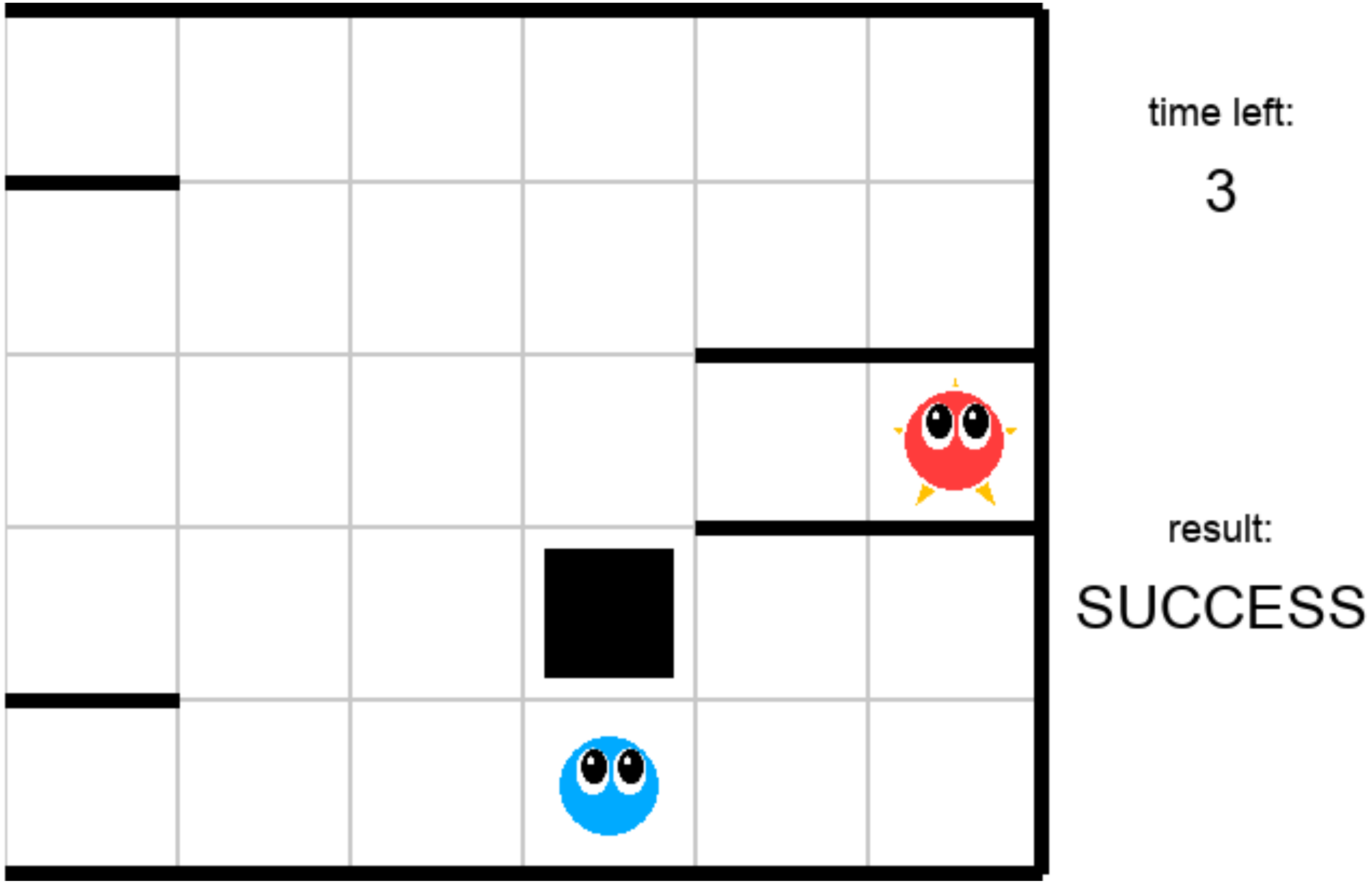
wants to help or hinder **red**



time left:

10

result:



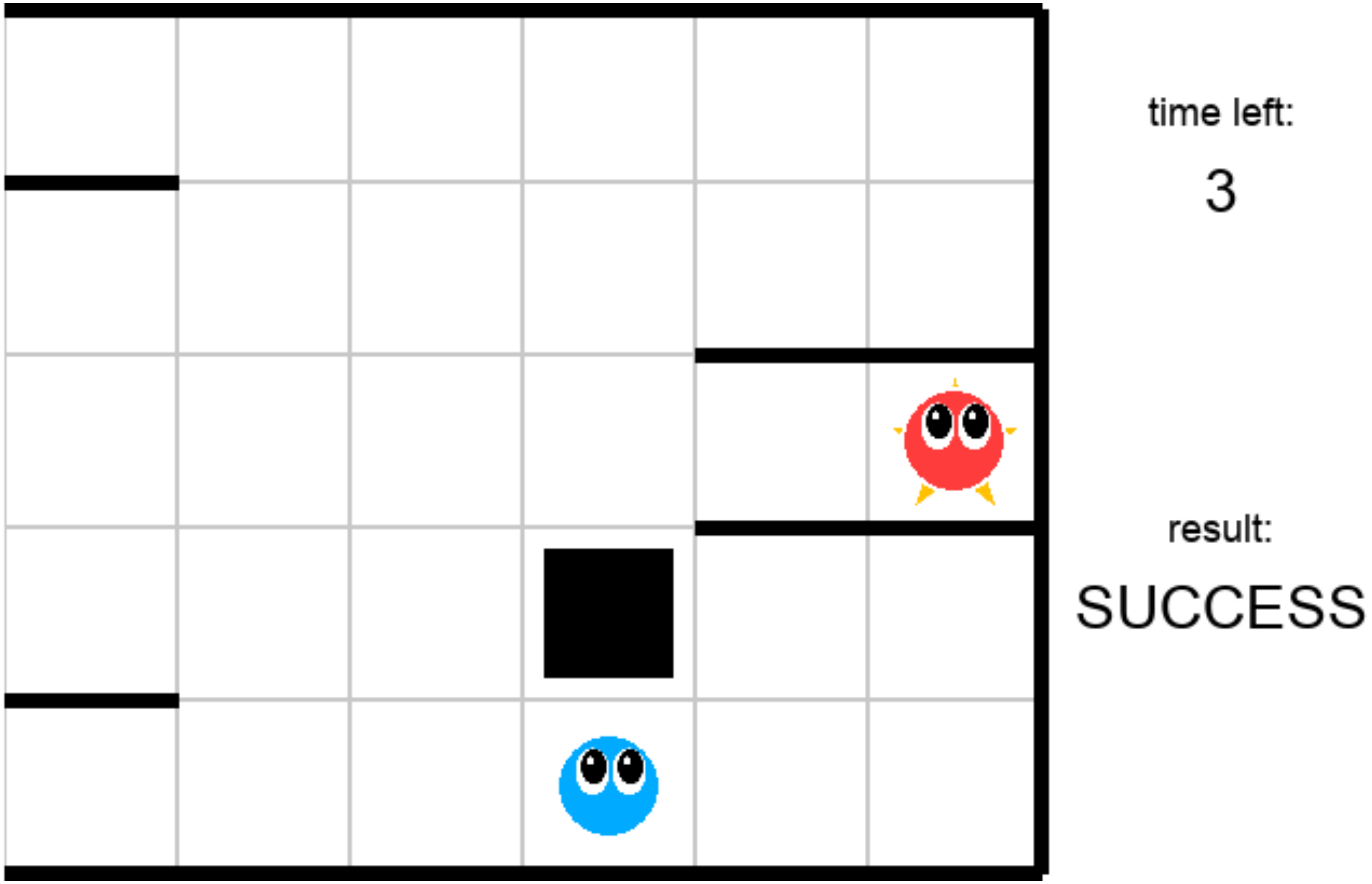
Cause

"The red player won because of the blue player."

don’t agree at all

agree very much

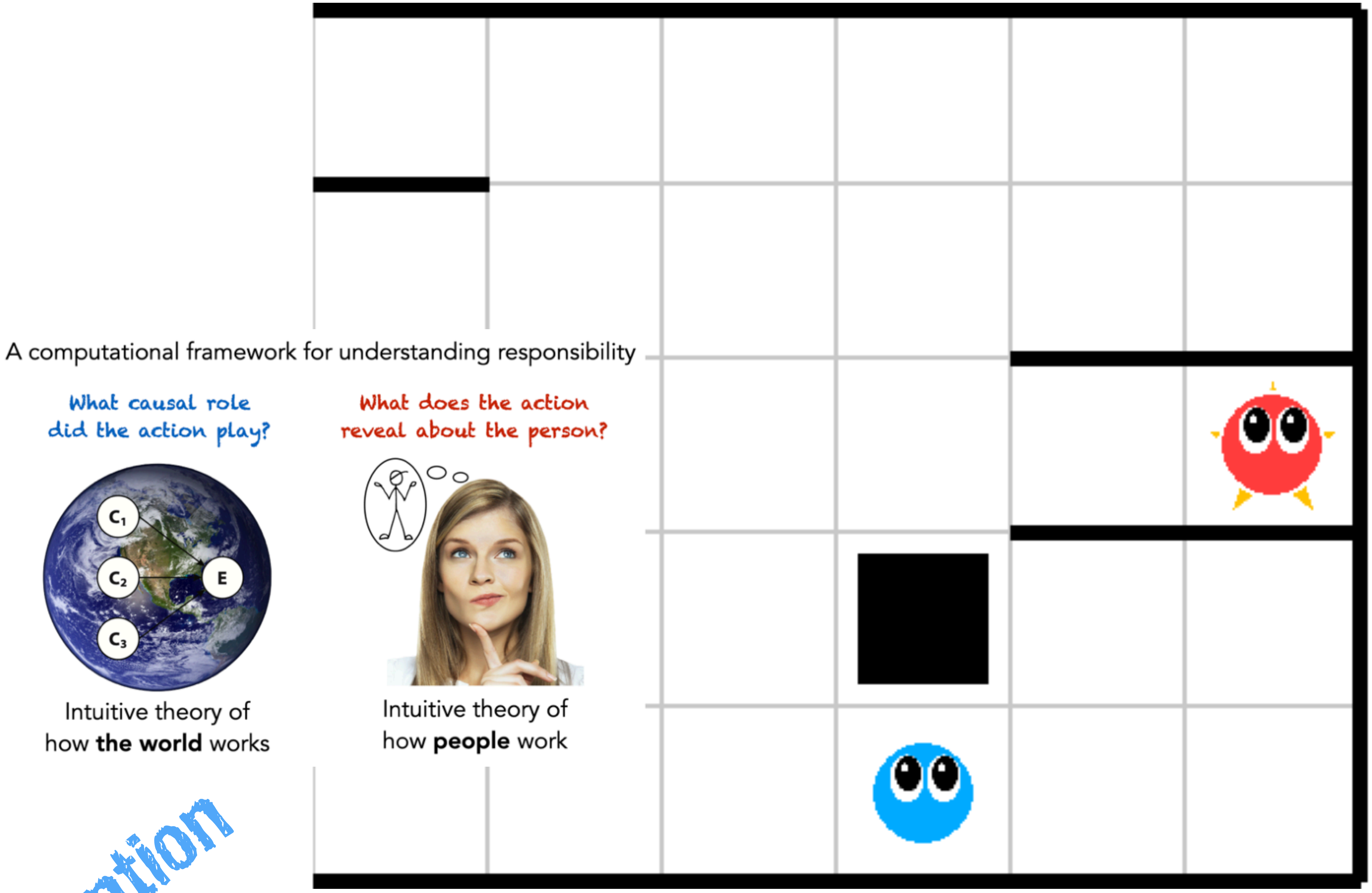
Counter factual



"The red player would still have succeeded if the blue player hadn't been there."

don’t agree at all

agree very much



A computational framework for understanding responsibility

What causal role did the action play?

What does the action reveal about the person?

Intuitive theory of how **the world** works

Intuitive theory of how **people** work

time left:

3

result:

SUCCESS

intention

What was the blue player intending to do?

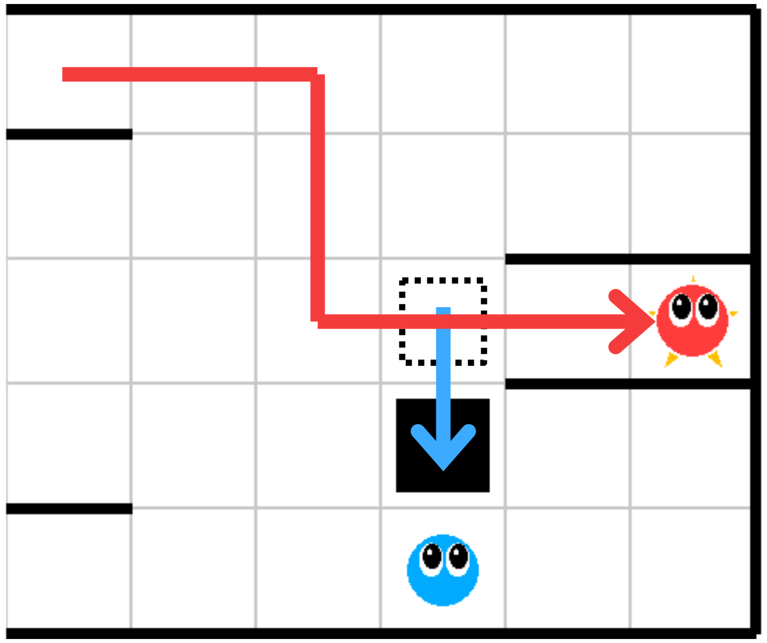
definitely hinder

definitely help

Counterfactual simulation model of causal judgment

actual situation:

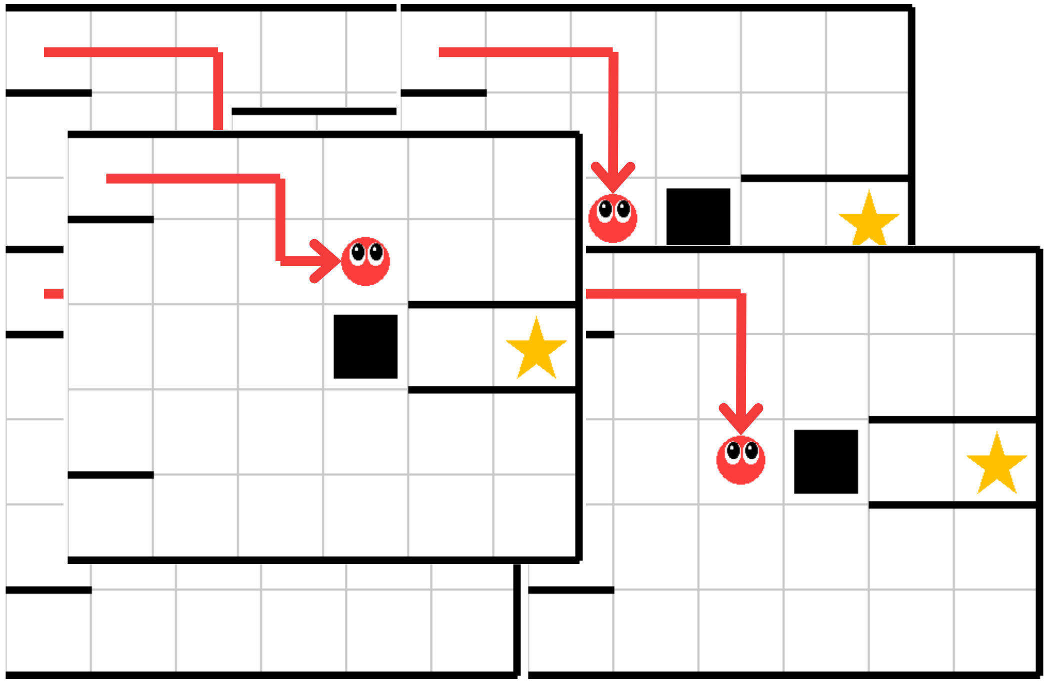
success



counterfactual simulations:

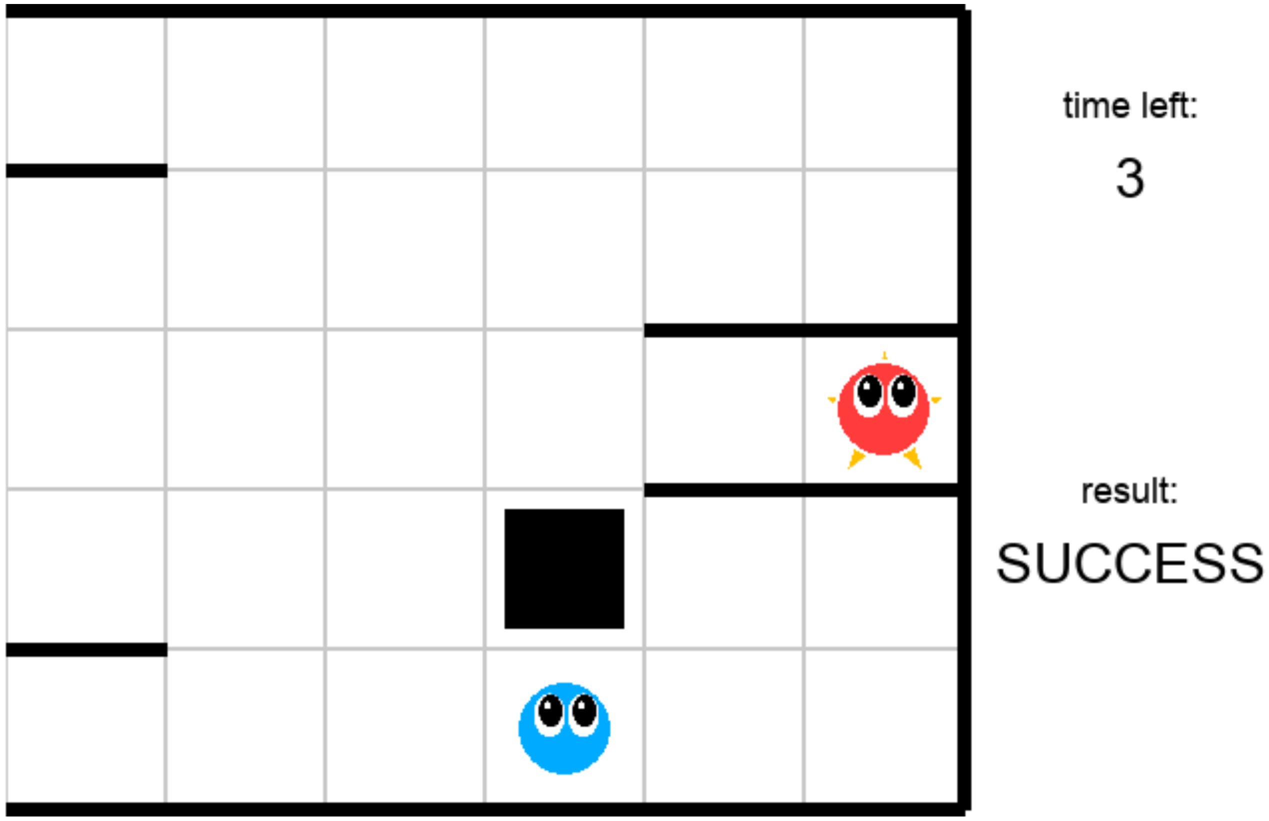
what would have happened

if **blue** hadn't been there



counterfactual outcome: 0% success

Intention inference model



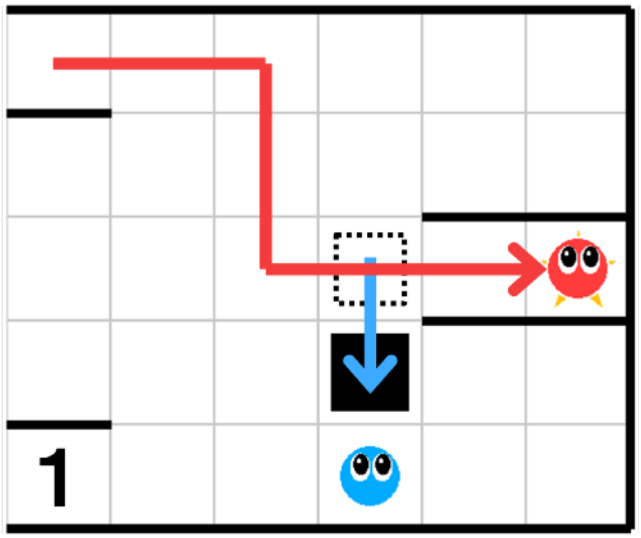
What was the blue player intending to do?

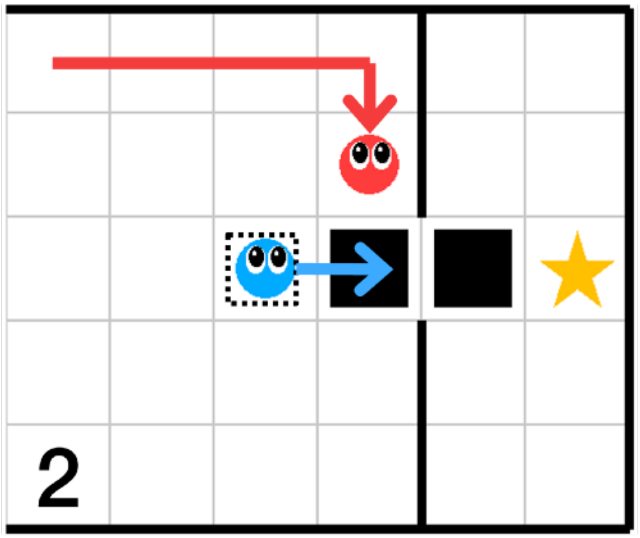
g_i = help or hinder agent j

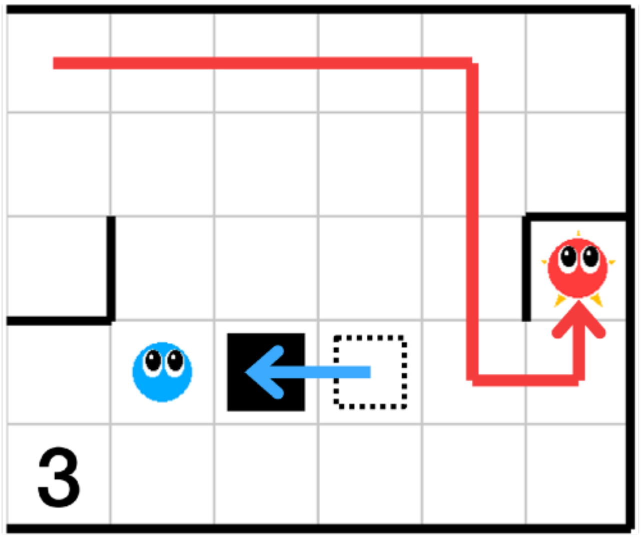
agent i learns policy through Monte Carlo tree search

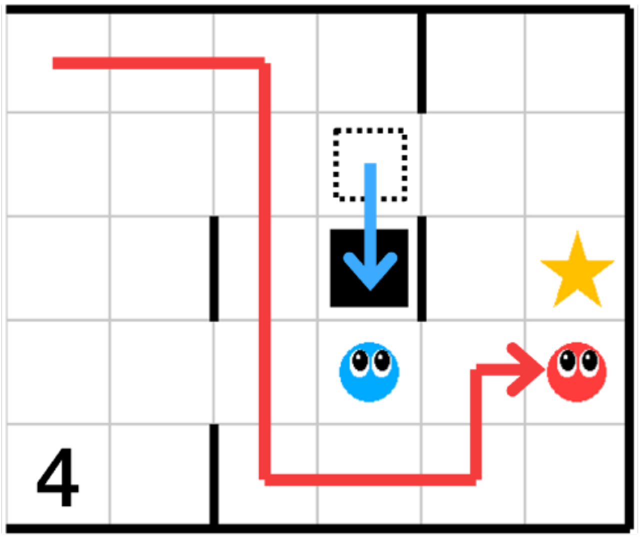
reward for each rollout depends on:

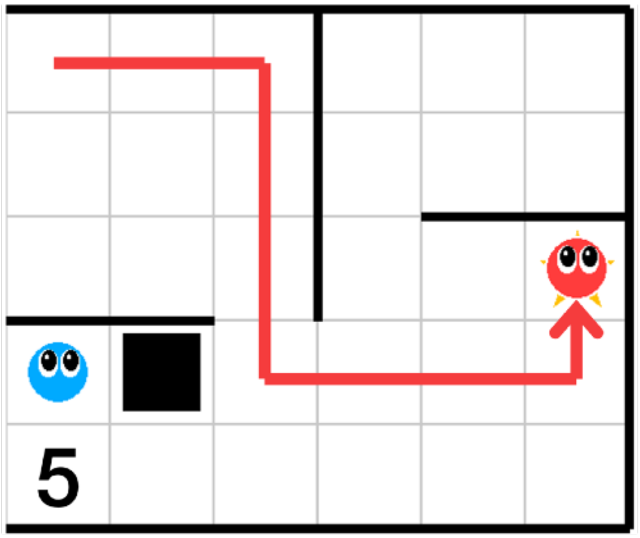
- agent i's utility
- agent j's utility
- number of available paths for agent j to goal


1

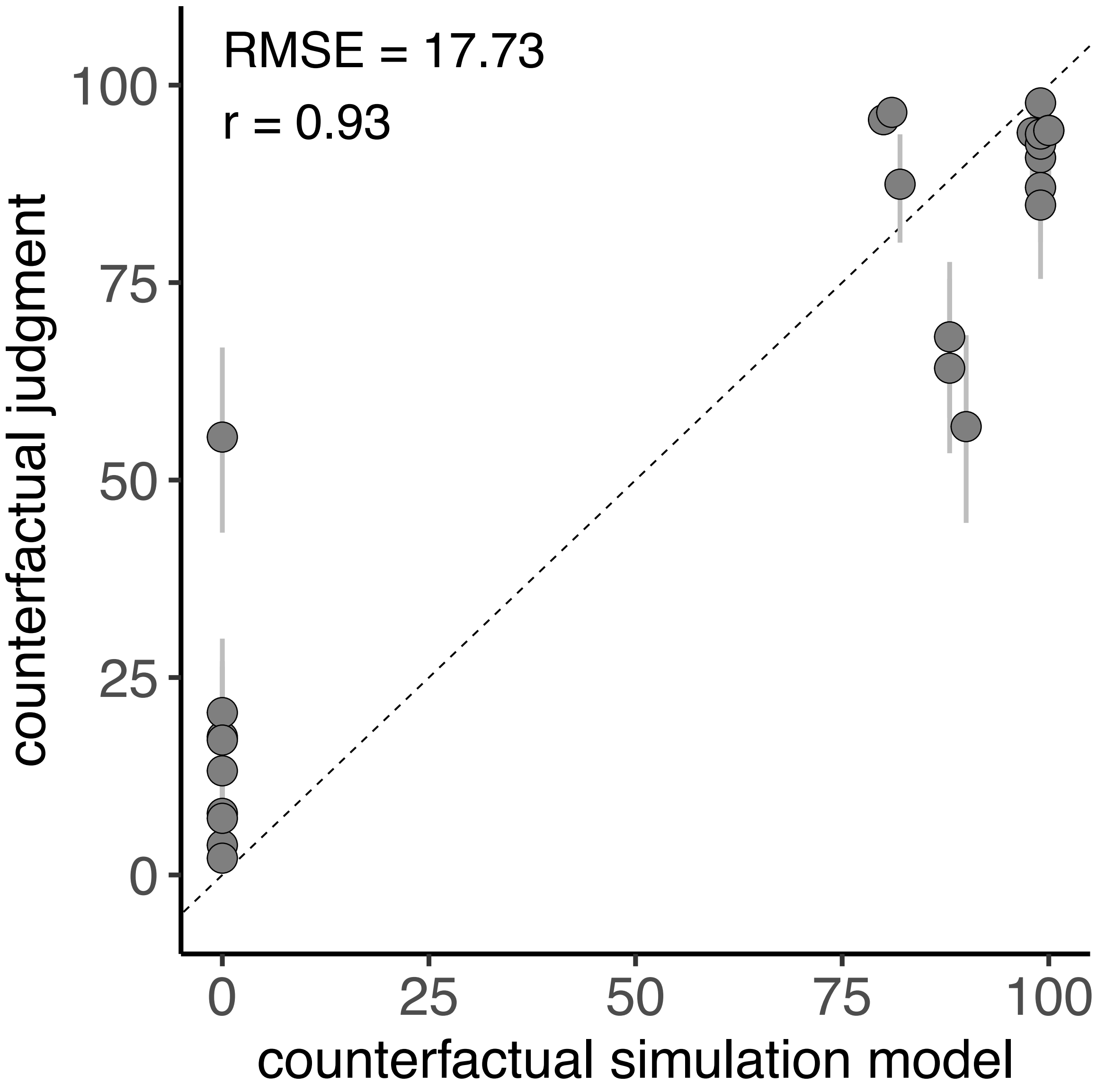

2

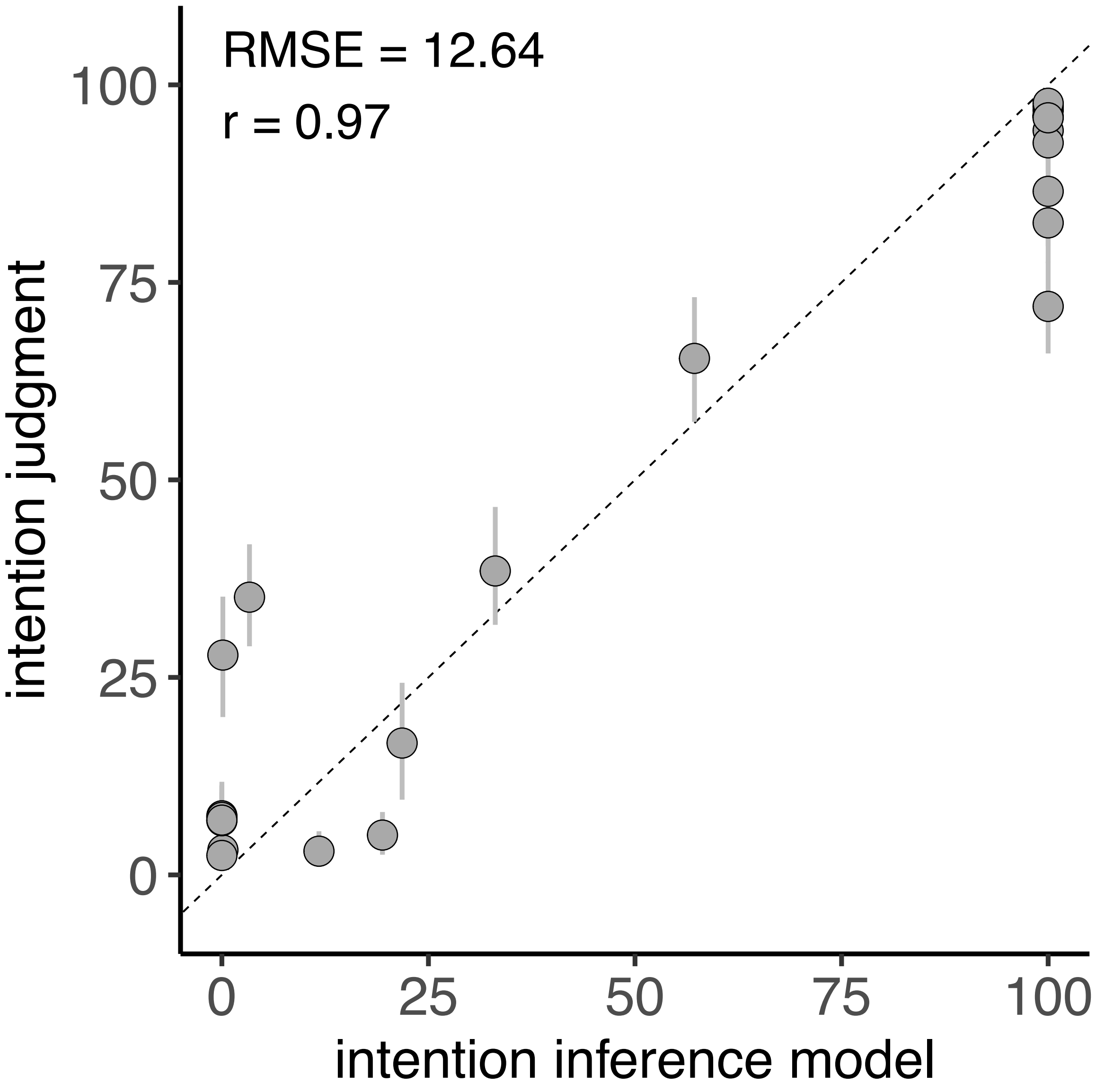

3


4


5

(n = 50 each)





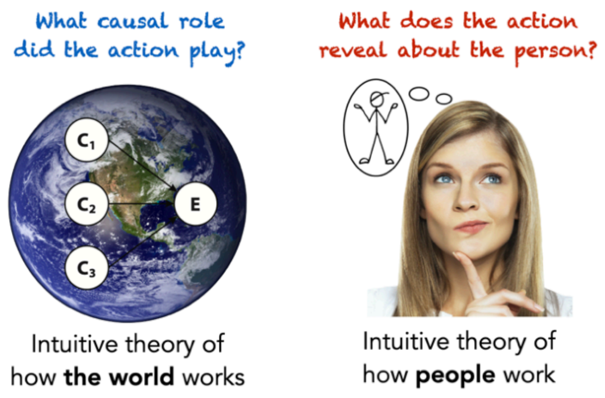
model captures much of the variance in counterfactual and intention judgments

OSF

(n = 50 each)

causal judgments

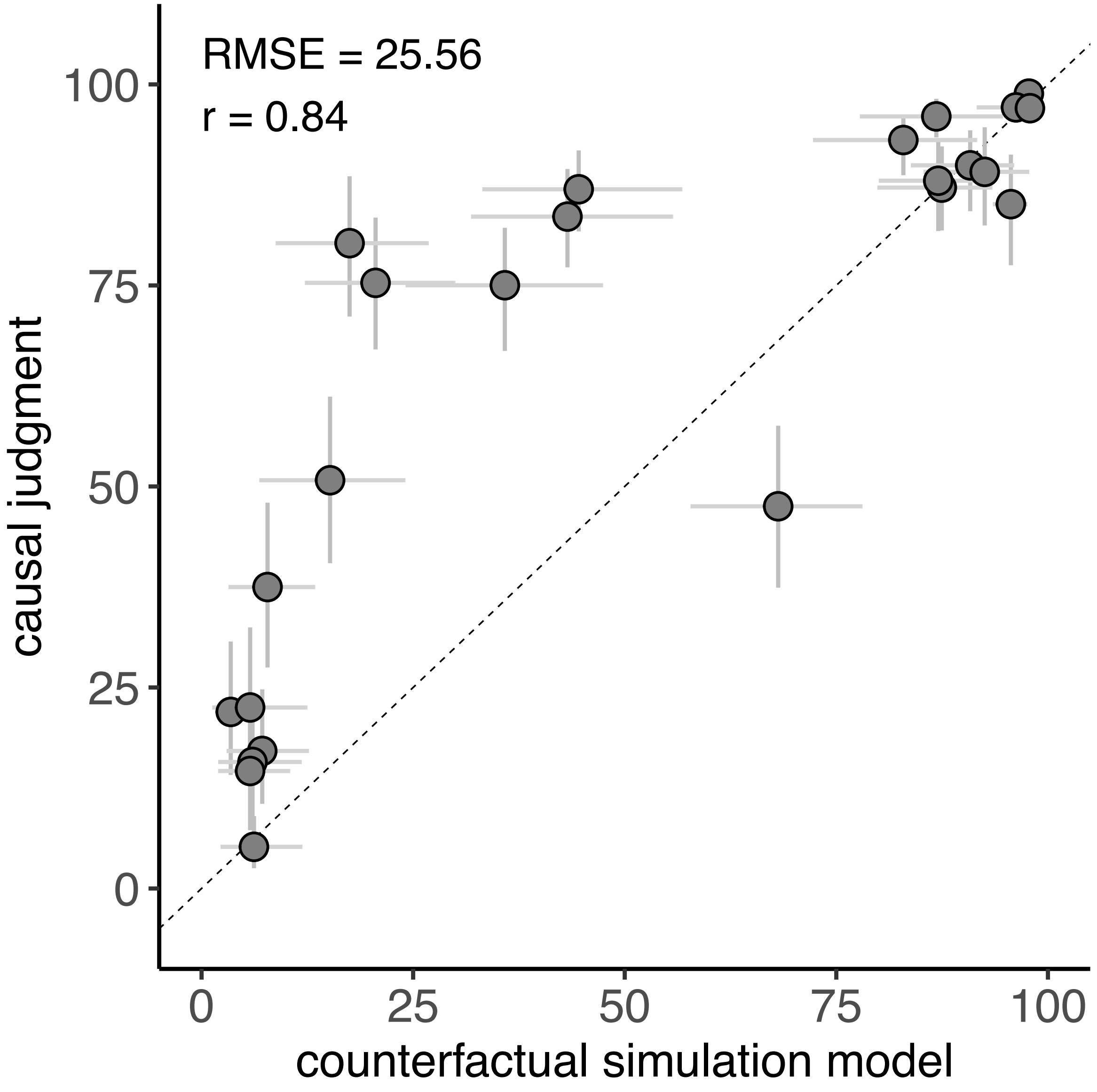
A computational framework for understanding responsibility

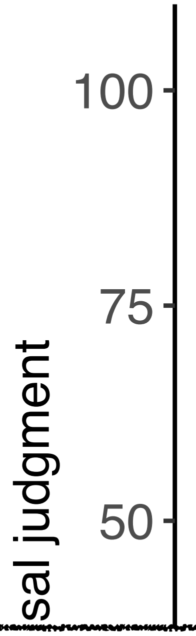

What causal role did the action play?

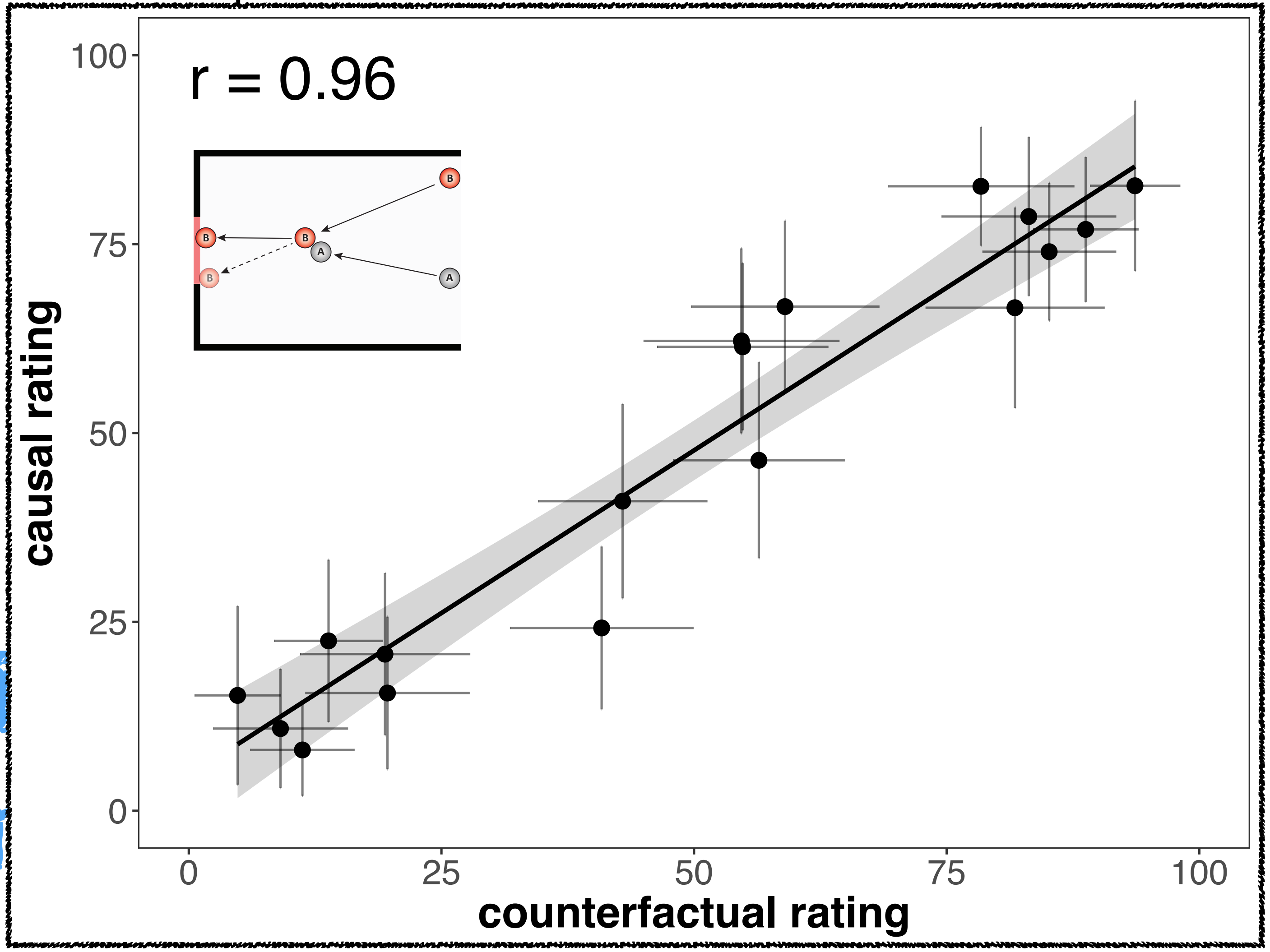
What does the action reveal about the person?

Intuitive theory of how **the world** works

Intuitive theory of how **people** work







doesn't look like this →

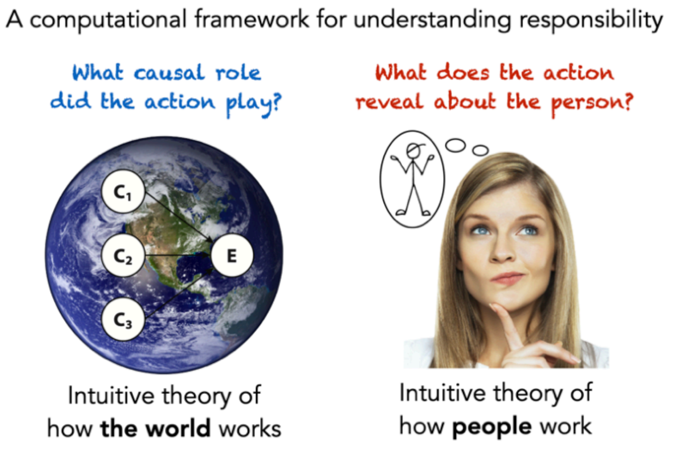
model that combi

simulation + intentio

blue's action made no difference

blue's intention was to hinder red

blue was judged to be responsible



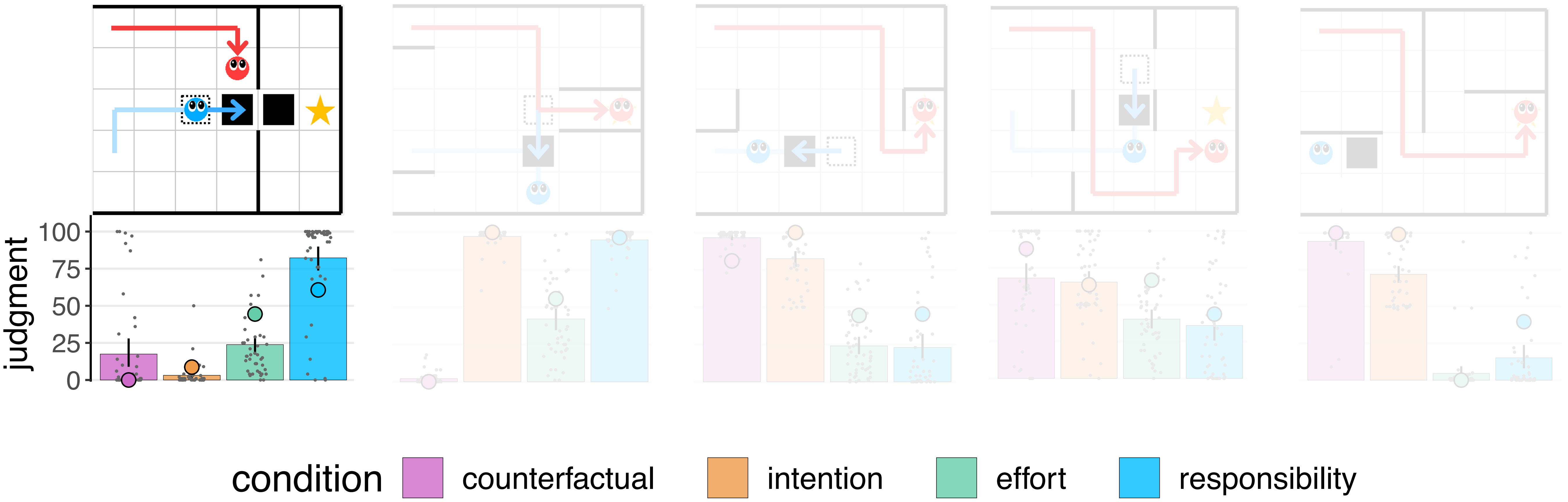
A computational framework for understanding responsibility

What causal role did the action play?

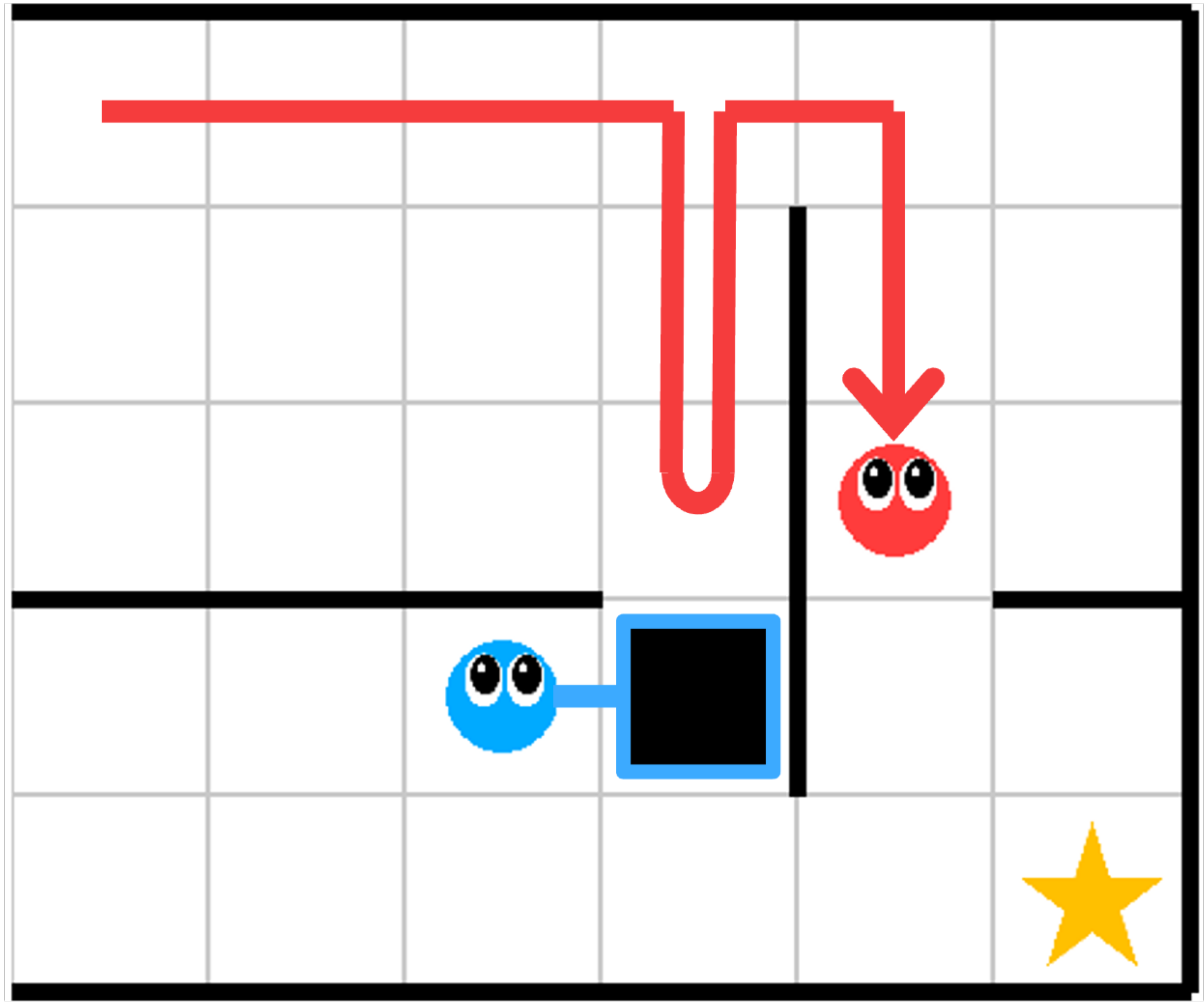
What does the action reveal about the person?

Intuitive theory of how **the world** works

Intuitive theory of how **people** work

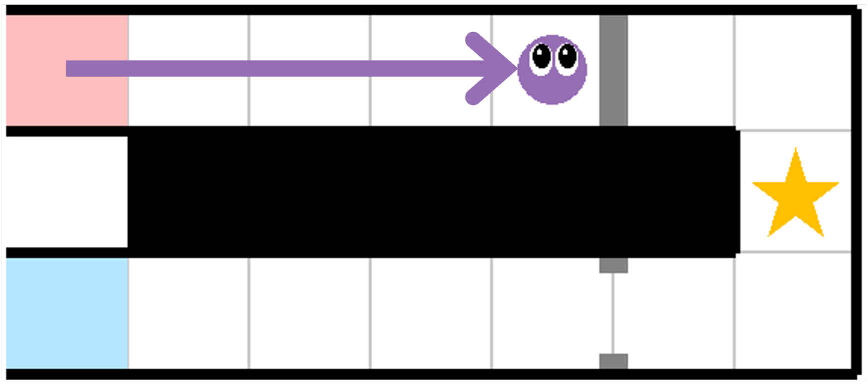


"BLUE tricked RED into thinking she was going to move the box to help her, but once RED was stuck on that side of the wall, BLUE left the box where it was."

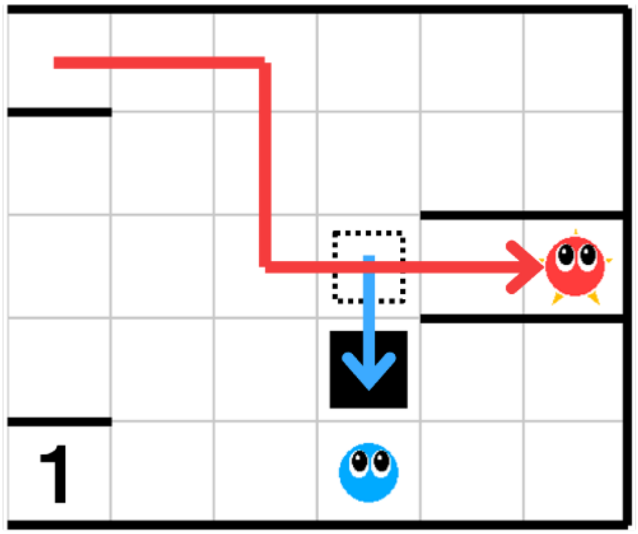


hindering doesn't require changing the physical world, it's enough to change someone's mind

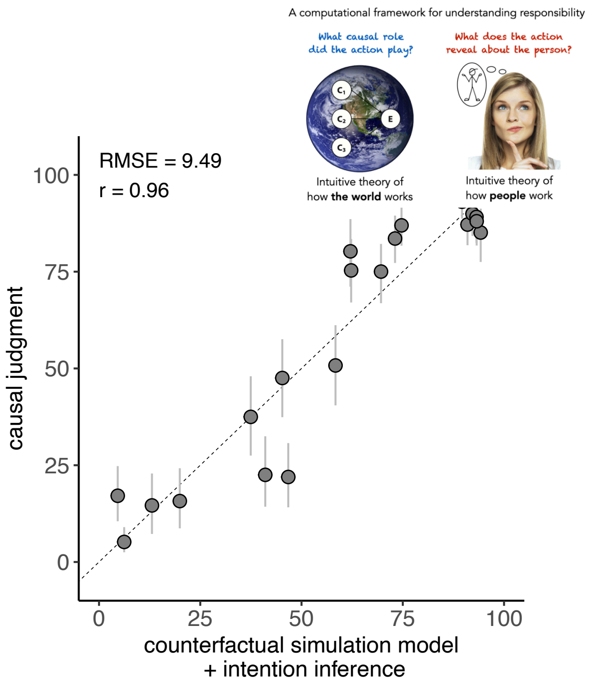
Counterfactual simulation model of causal judgment



- people give causal explanations about agents' actions by **simulating counterfactuals**



- judging whether someone **helped or hindered** requires counterfactual simulation



- explanations in social settings are sensitive to the agent's **causal role** and their **inferred mental states**

Conclusion

- we build rich mental **models of the world**
- by imagining interventions and running **mental simulations**, we can compute counterfactuals which are critical for giving causal explanations
- the counterfactual simulation model captures causal judgments about **physical events** and **social events**

Gerstenberg & Tenenbaum (2017) Intuitive Theories. *Oxford Handbook of Causal Reasoning*

Goodman, Tenenbaum, & Gerstenberg (2015) Concepts in a probabilistic language of thought. *The Conceptual Mind: New Directions in the Study of Concepts*

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Wu, Sridhar, & Gerstenberg (2022) That was close! A counterfactual simulation model of causal judgments about decisions. *CogSci Proceedings*

Thanks!

Josh Tenenbaum

David Lagnado

Noah Goodman

Matt Peterson

Sarah Wu

Shruti Sridhar

@tobigerstenberg

http://cicl.stanford.edu

Conclusion

- we build rich mental **models of the world**

- by imagining interventions and running **mental simulations**, we can compute counterfactuals which are critical for giving causal explanations

- the counterfactual simulation model captures causal judgments about **physical events** and **social events**

Gerstenberg & Tenenbaum (2017) Intuitive Theories. *Oxford Handbook of Causal Reasoning*

Goodman, Tenenbaum, & Gerstenberg (2015) Concepts in a probabilistic language of thought. *The Conceptual Mind: New Directions in the Study of Concepts*

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Wu, Sridhar, & Gerstenberg (2022) That was close! A counterfactual simulation model of causal judgments about decisions. *CogSci Proceedings*