Ethics and Limitations

By Helena Vasconcelos and Carolyn Zou

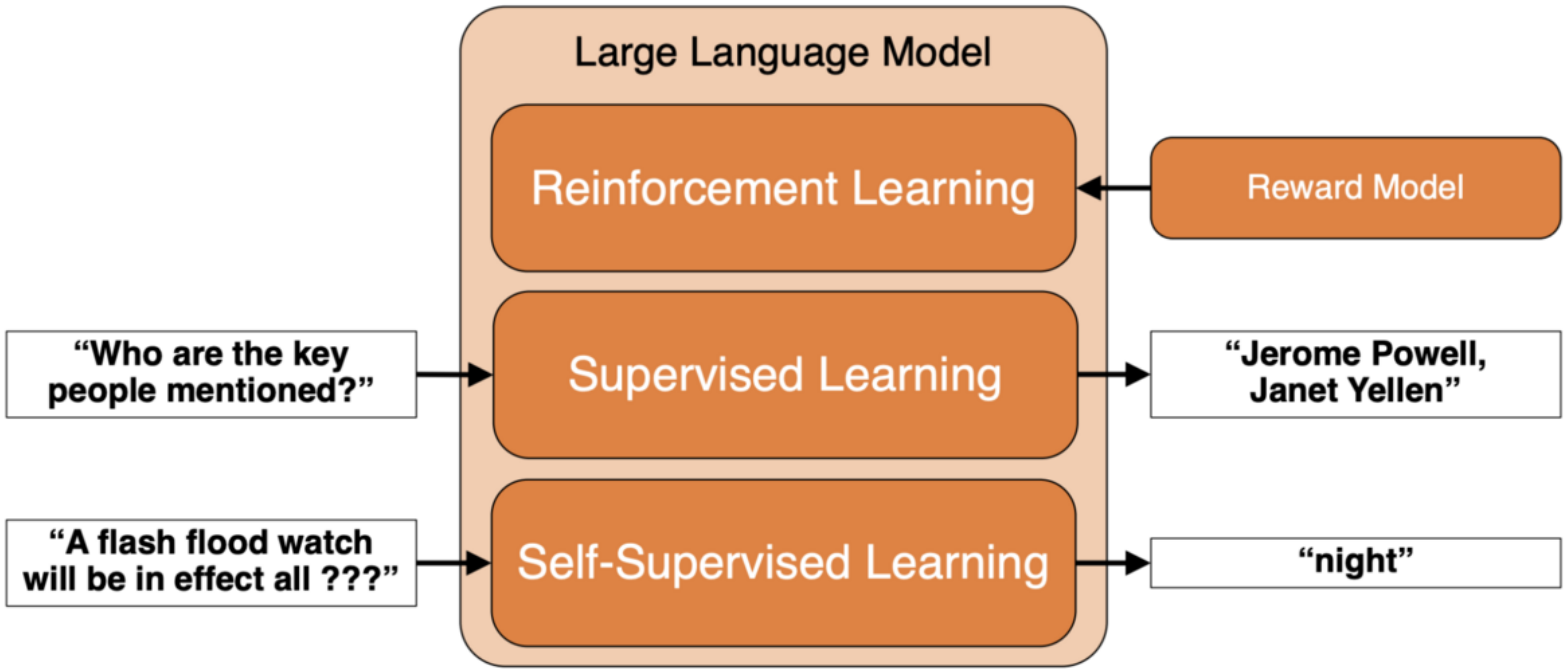
Lecture Roadmap:

LLM Training 🤔	Training Data 📈	Running Inference 🏃	Validation ✔	Reliance 👀
How do the ways (e.g., RLHF) in which models are trained affect the behaviors of agents?	What have these models learned? From where? How does this limit the accuracy of our agents?	How does the stochasticity and memorization of models affect accuracy? How do the architectures of the agents affect things?	Given the outputs of simulations, how might we validate them? How do we know what to trust?	How much trust should we put into the results of simulations? What happens if we put too much trust?

Lecture Roadmap:

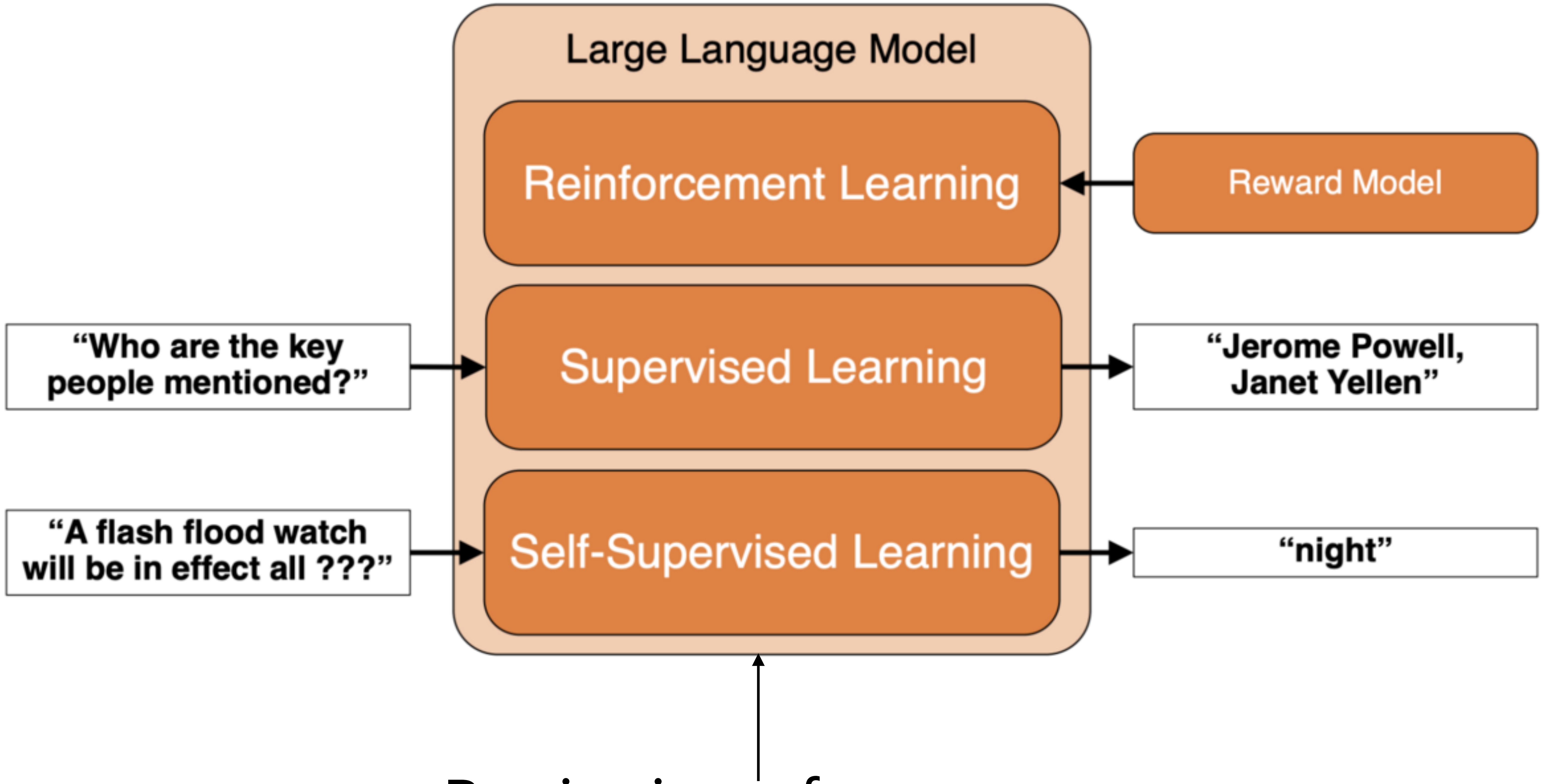
LLM Training 🤔	Training Data 📈	Running Inference 🏃	Validation ✔	Reliance 👀
How do the ways (e.g., RLHF) in which models are trained affect the behaviors of agents?	What have these models learned? From where? How does this limit the accuracy of our agents?	How does the stochasticity and memorization of models affect accuracy? How do the architectures of the agents affect things?	Given the outputs of simulations, how might we validate them? How do we know what to trust?	How much trust should we put into the results of simulations? What happens if we put too much trust?

How Generative AI is Made



Credit Stephen Bach, former post-doc at Stanford, now at Brown CS

How Generative AI is Made



Beginning of process

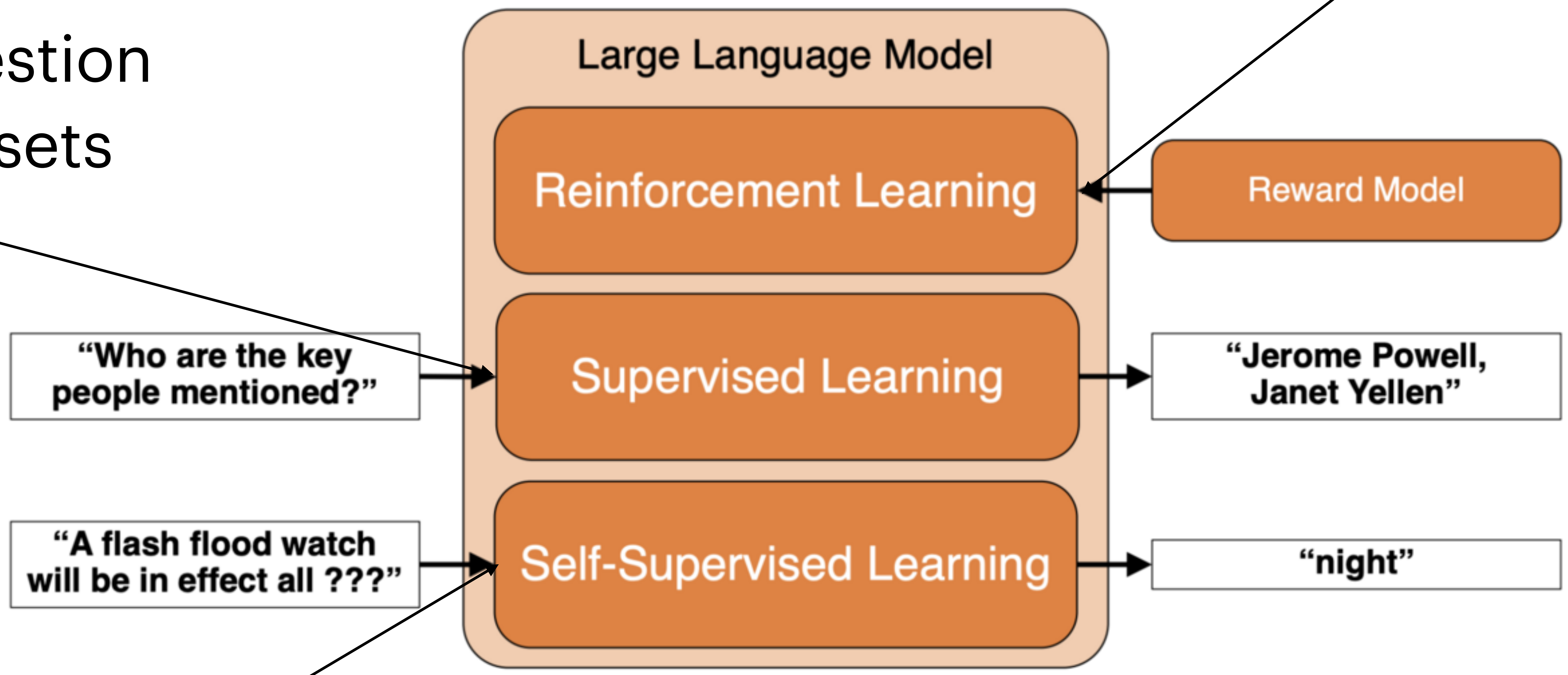
Credit Stephen Bach, former post-doc at Stanford, now at Brown CS

*we'll talk a bit more about data!

Preference Data

How Generative AI is Made

Informative Question Answering Datasets



Giant corpus from web

Credit Stephen Bach, former post-doc at Stanford, now at Brown CS

The broader point is: these models are *not* optimized to act like people.

The broader point is: these models are *not* optimized to act like people.

* some researchers are trying to retrain models such that they are trained to predict behavior, but this is still early work!

“Humanlike behaviors”

- Next token prediction is somewhat unintuitive
- So in order for LLMs to be useful products, their behaviors should be more recognizable to the average person
- The jump from gpt-3 to ChatGPT: instruction tuning
  - completion vs chat
  - The system is humanlike*

  * but always follows instructions

“Humanlike behaviors”

- But we don’t want to “talk” to any random person
- Our assistant should be knowledgeable, friendly, helpful, etc.
- Hence, RLHF

- The system is humanlike*

  * but always follows instructions, always knows the “answer”, is friendly…

Lecture Roadmap:

LLM Training 🤔	Training Data 📈	Running Inference 🏃‍♀️	Validation ✔	Reliance 👀
How do the ways (e.g., RLHF) in which models are trained affect the behaviors of agents?	What have these models learned? From where? How does this limit the accuracy of our agents?	How does the stochasticity and memorization of models affect accuracy? How do the architectures of the agents affect things?	Given the outputs of simulations, how might we validate them? How do we know what to trust?	How much trust should we put into the results of simulations? What happens if we put too much trust?

What kind of data do models like ChatGPT use?

- Data scraped from the web (e.g., Wikipedia, Reddit)
- Q&A, informational data
- RLHF data (e.g., paired rankings on quality of certain responses)

These data sources limit what we can do...

- For some tasks, simulations might be more appropriate (e.g., tasks that emulate online dynamics or are primarily “knowledge based"), since that is closer to the training data
- Other tasks (e.g., tasks that require physical dynamics) do not translate well from the LLM paradigm

Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents

JUNKAI LI[†#], SIYU WANG[†], MENG ZHANG[†], WEITAO LI[†#], YUNGHWEI LAI[†], XINHUI KANG[†#], WEIZHI MA[†], and YANG LIU[#†]

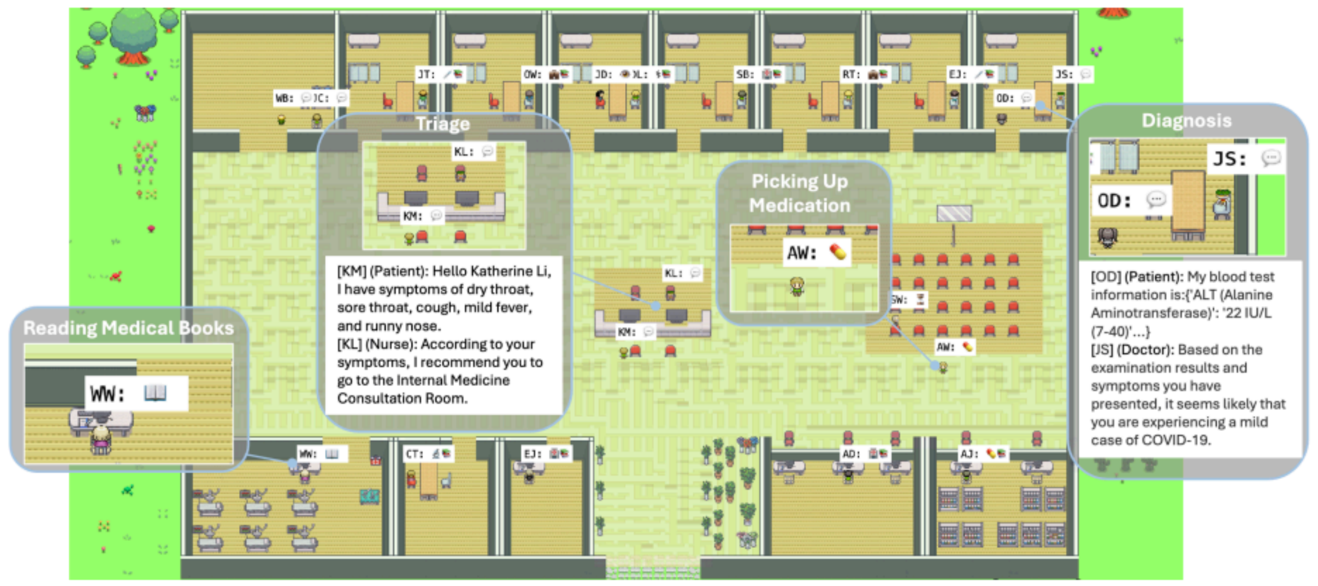


Fig. 1. An overview of Agent Hospital. It is a simulacrum of hospital in which patients, nurses, and doctors are autonomous agents powered by large language models. Agent Hospital simulates the whole closed cycle of treating a patient’s illness: disease onset, triage, registration, consultation, medical examination, diagnosis, medicine dispensary, convalescence, and post-hospital follow-up visit. An interesting finding is that the doctor agents can keep improving treatment performance over time without manually labeled data, both in simulation and real-world evaluations.

These data sources limit what we can do...

Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models

Myra Cheng
Stanford University
myra@cs.stanford.edu

Esin Durmus
Stanford University

Dan Jurafsky
Stanford University

Abstract

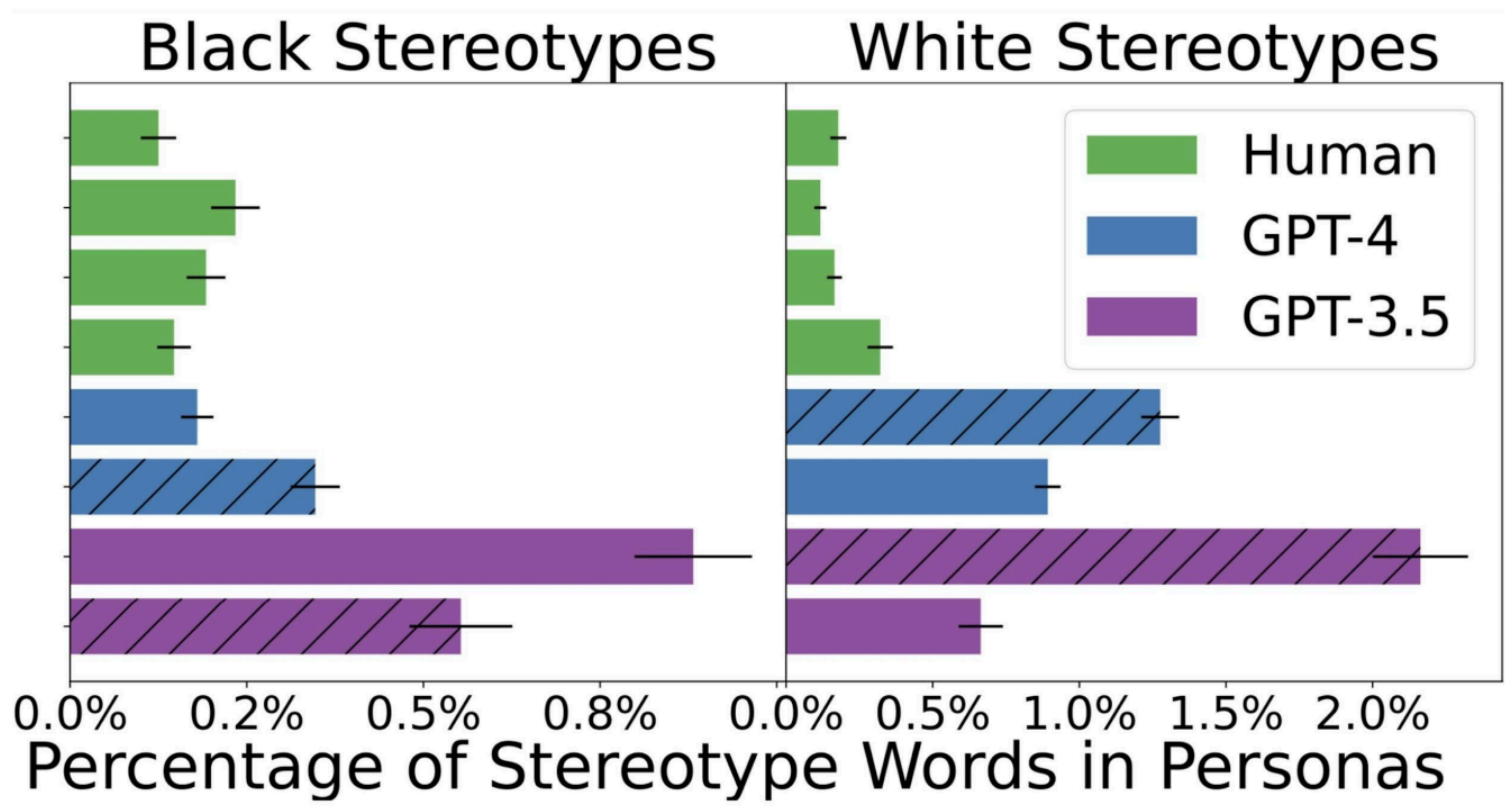
To recognize and mitigate harms from large language models (LLMs), we need to understand the prevalence and nuances of stereotypes in LLM outputs. Toward this end, we present **Marked Personas**, a prompt-based method to measure stereotypes in LLMs for intersectional demographic groups without any lexicon or data labeling. Grounded in the sociolinguistic concept of *markedness* (which characterizes explicitly linguistically marked categories versus unmarked defaults), our proposed method is twofold: 1) prompting an LLM to generate personas, i.e., natural language descriptions, of the target demographic group alongside personas of unmarked, default groups; 2) identifying the words that significantly distinguish personas of the target group from corresponding unmarked ones. We find that the portrayals generated by GPT-3.5 and GPT-4 contain higher rates of racial stereotypes than human-written portrayals using the same prompts. The words

As I look in the mirror, I see my **rich**, **melanin**-infused skin glowing softly. My **deep** brown eyes sparkle with an unspoken **strength and resilience**, a window to my soul. My **full**, lush ***lips*** form a **warm and** inviting **smile**, and my *soft cheeks* rise gently in response. My hair, a riot **of textured coils**, frames my face in a **gravity**-defying halo. It dances to its own beat, wild **and** free, just like me. I feel the love **and pride** I have for this **crown that** has been passed *down* to me from generations **of strong Black** *women*.

Table 1: Example of GPT-4-generated persona of a Black woman. **Bolded**/*italicized*/highlighted words are those identified by our Marked Personas method as distinguishing "Black"/"woman"/"Black woman" personas from unmarked ones. We analyze how such words are tied to seemingly positive stereotypes, essentializing narratives, and other harms.

Nadeem et al., 2021). They also have a trade-off between 1) characterizing a fixed set of stereotypes

Generated personas contain more stereotypes



What was in the reading?

Large language models should not replace human participants because they can misportray and flatten identity groups

Angelina Wang[1], Jamie Morgenstern[2], John P. Dickerson[3,4]

[1]Computer Science, Stanford University, Palo Alto, CA, USA.
[2]Computer Science & Engineering, University of Washington, Seattle, WA, USA.
[3]Computer Science, University of Maryland, College Park, MD, USA.
[4]Arthur, New York City, NY, USA.

Contributing authors: angelina.wang@stanford.edu; jamiemmt@cs.washington.edu; john@arthur.ai;

Abstract

Large language models (LLMs) are increasing in capability and popularity, propelling their application in new domains—including as replacements for human participants in computational social science, user testing, annotation tasks, and more. In many settings, researchers seek to distribute their surveys to a sample of participants that are representative of the underlying human population of interest. This means in order to be a suitable replacement, LLMs will need to be able to capture the influence of positionality (i.e., relevance of social identities like gender and race). However, we show that there are two inherent limitations in the way current LLMs are trained that prevent this. We argue analytically for why LLMs are likely to both *misportray* and *flatten* the representations of demographic groups, then empirically show this on 4 LLMs through a series of human studies with 3200 participants across 16 demographic identities. We also discuss a third limitation about how identity prompts can essentialize identities. Throughout, we connect each limitation to a pernicious history that explains why it is harmful for marginalized demographic groups. Overall, we urge caution in use cases where LLMs are intended to replace human participants whose identities are relevant to the task at hand. At the same time, in cases where the goal is to supplement rather than replace (e.g., pilot studies), we provide inference-time techniques that we empirically demonstrate do reduce, but do not remove, these harms.

Whose Opinions Do Language Models Reflect?

Shibani Santurkar, Stanford, shibani@stanford.edu
Esin Durmus, Stanford, esindurmus@cs.stanford.edu
Faisal Ladhak, Columbia University, faisal@cs.columbia.edu
Cinoo Lee, Stanford, cinoolee@stanford.edu
Percy Liang, Stanford, pliang@cs.stanford.edu
Tatsunori Hashimoto, Stanford, thashim@stanford.edu

Abstract

Language models (LMs) are increasingly being used in open-ended contexts, where the opinions reflected by LMs in response to subjective queries can have a profound impact, both on user satisfaction, as well as shaping the views of society at large. In this work, we put forth a quantitative framework to investigate the opinions reflected by LMs – by leveraging high-quality public opinion polls and their associated human responses. Using this framework, we create OpinionQA, a new dataset for evaluating the alignment of LM opinions with those of 60 US demographic groups over topics ranging from abortion to automation. Across topics, we find substantial misalignment between the views reflected by current LMs and those of US demographic groups: on par with the Democrat-Republican divide on climate change. Notably, this misalignment persists even after explicitly steering the LMs towards particular demographic groups. Our analysis not only confirms prior observations about the left-leaning tendencies of some human feedback-tuned LMs, but also surfaces groups whose opinions are poorly reflected by current LMs (e.g., 65+ and widowed individuals). Our code and data are available at `https://github.com/tatsu-lab/opinions_qa`.

These data sources limit what we can do…

Systematic Biases in LLM Simulations of Debates

Amir Taubenfeld[12*] **Yaniv Dover**[34] **Roi Reichart**[5] **Ariel Goldstein**[236]

*Corresponding Author: amirt@google.com

[1]The Hebrew University of Jerusalem, School of Computer Science and Engineering
[2]Google Research
[3]The Hebrew University Business School, Jerusalem, Israel
[4]Federmann Center for the Study of Rationality, Hebrew University, Jerusalem, Israel
[5]Faculty of Data and Decision Sciences, Technion
[6]Department of Cognitive and Brain Sciences, Hebrew University, Jerusalem, Israel

Abstract

The emergence of Large Language Models (LLMs), has opened exciting possibilities for constructing computational simulations designed to replicate human behavior accurately. Current research suggests that LLM-based agents become increasingly human-like in their performance, sparking interest in using these AI agents as substitutes for human participants in behavioral studies. However, LLMs are complex statistical learners without straightforward deductive rules, making them prone to unexpected behaviors. Hence, it is crucial to study and pinpoint the key behavioral distinctions be-

aim to accurately replicate human behavior (Park et al., 2023; Qian et al., 2023). Current research suggests that LLM-based agents become increasingly human-like in their performance and that they possess the remarkable ability to seamlessly adopt personas of different characters (Shanahan et al., 2023; Argyle et al., 2023). The typical paradigm for such simulations involves selecting an LLM, such as the widely used ChatGPT (Milmo, 2023), as a base model and crafting individual agents' identities through natural language prompts. For instance, by prepending the prompt, "John Lin is a pharmacy shopkeeper," to an agent's context, the

These data sources limit what we can do...

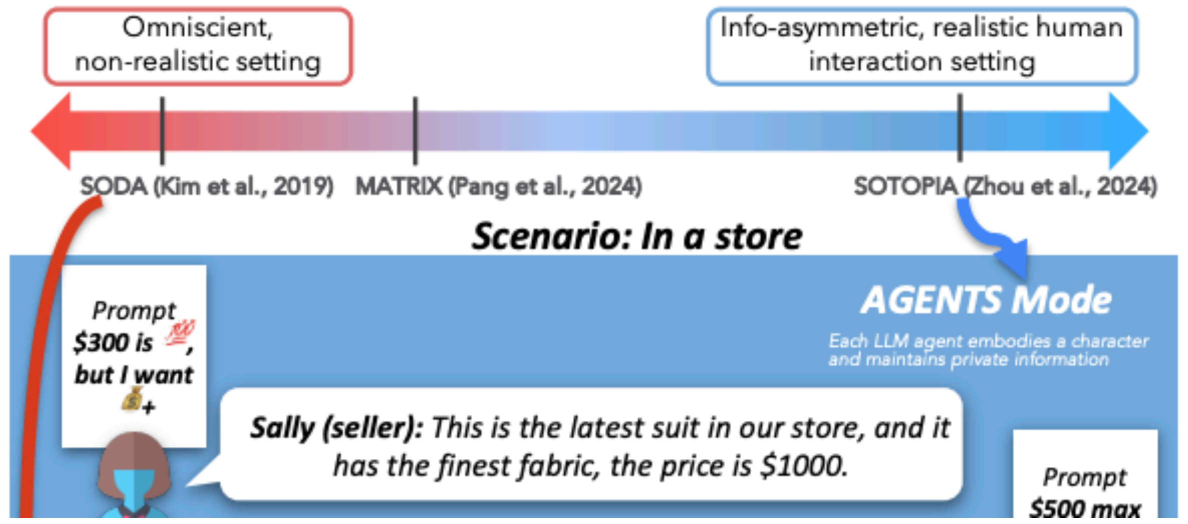
Is this the real life? Is this just fantasy?
The Misleading Success of Simulating Social Interactions With LLMs

Xuhui Zhou[♡] Zhe Su[♡] Tiwalayo Eisape[♠]
Hyunwoo Kim[♣] Maarten Sap[♡♣]

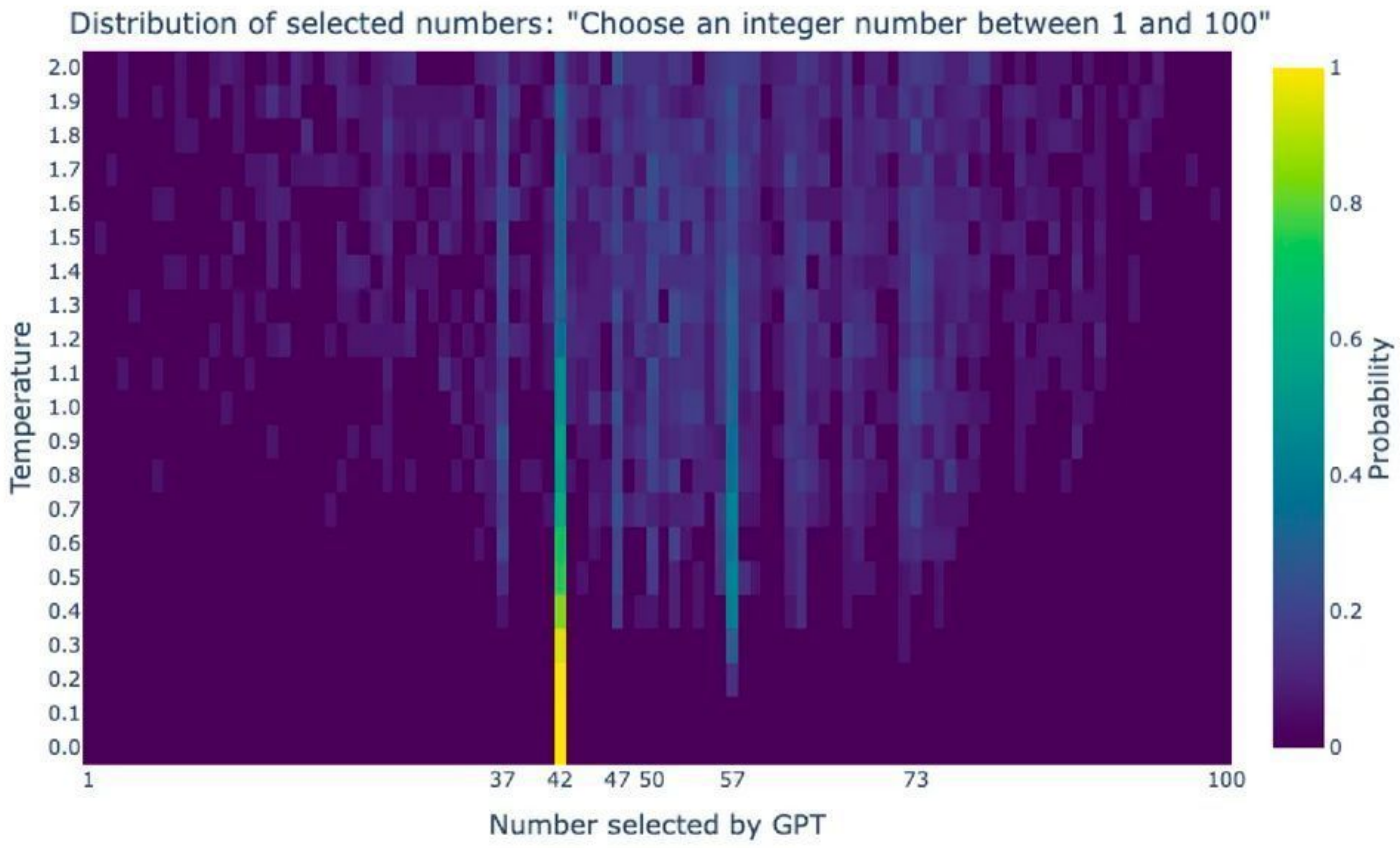
[♡]Carnegie Mellon University [♠]Massachusetts Institute of Technology [♣]Allen Institute for AI
xuhuiz@cs.cmu.edu agscr.sotopia.world

Abstract

Recent advances in large language models (LLM) have enabled richer social simulations, allowing for the study of various social phenomena. However, most recent work has used a more omniscient perspective on these simulations (e.g., single LLM to generate all inter-



These data sources limit what we can do...



AI Research
by Leniolabs_

Source: ChatGPT prompted 1000 times with
"Choose an integer number between 1 and 100"

What other data sources do you think would affect the realism of agents?

Lecture Roadmap:

LLM Training 🤔	Training Data 📈	Running Inference 🏃‍♀️	Validation ✔	Reliance 👀
How do the ways (e.g., RLHF) in which models are trained affect the behaviors of agents?	What have these models learned? From where? How does this limit the accuracy of our agents?	How does the stochasticity and memorization of models affect accuracy? How do the architectures of the agents affect things?	Given the outputs of simulations, how might we validate them? How do we know what to trust?	How much trust should we put into the results of simulations? What happens if we put too much trust?

Replication?

Replication crisis?

PLOS MEDICINE

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Simulation robustness

- When creating agents for simulations, it's tempting to use human behavior metaphors for sensemaking
- But three behaviors unique to language models make this a bad idea:
    - Prompt sensitivity
    - Stochasticity
    - Memorization
- Two approaches to identify and measure: **perturb** and **iterate**

Perturb

Dimension	Probe
Protocol	Expand study conditions trivially
Language	Rewrite prompts while preserving semantics
Settings	Iterate over hyperparameters, model versions
Format	Alter input/output formatting, digits, newlines
Strategy	With(out) chain-of-thought, preamble elements

Iterate

- Prompt draws from a (hidden) population
- Many draws produce a (simulated) sample
- Many samples produce a sampling distribution

Prompt sensitivity

BEHAVIORAL ECONOMICS

Avoiding overhead aversion in charity

Uri Gneezy,[1,2*] Elizabeth A. Keenan,[1] Ayelet Gneezy[1]

Donors tend to avoid charities that dedicate a high percentage of expenses to administrative and fundraising costs, limiting the ability of nonprofits to be effective. We propose a solution to this problem: Use donations from major philanthropists to cover overhead expenses and offer potential donors an overhead-free donation opportunity. A laboratory experiment testing this solution confirms that donations decrease when overhead increases, but only when donors pay for overhead themselves. In a field experiment with 40,000 potential donors, we compared the overhead-free solution with other common uses of initial donations. Consistent with prior research, informing donors that seed money has already been raised increases donations, as does a $1:$1 matching campaign. Our main result, however, clearly shows that informing potential donors that overhead costs are covered by an initial donation significantly increases the donation rate by 80% (or 94%) and total donations by 75% (or 89%) compared with the seed (or matching) approach.

[Gneezy et al.. Avoiding overhead aversion in charity. *Science* **346**,632-635 (2014). DOI:10.1126/science.1253932]

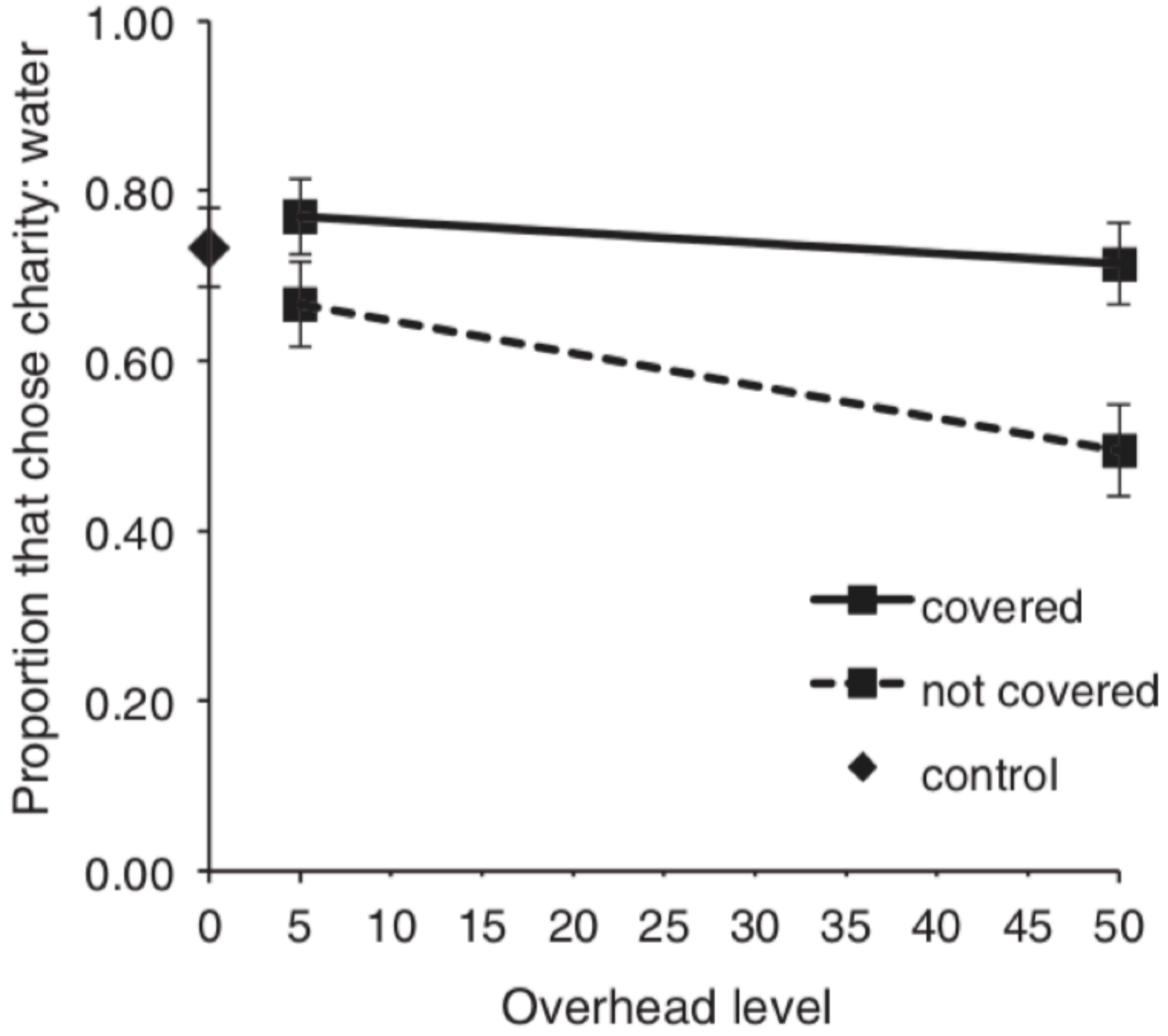
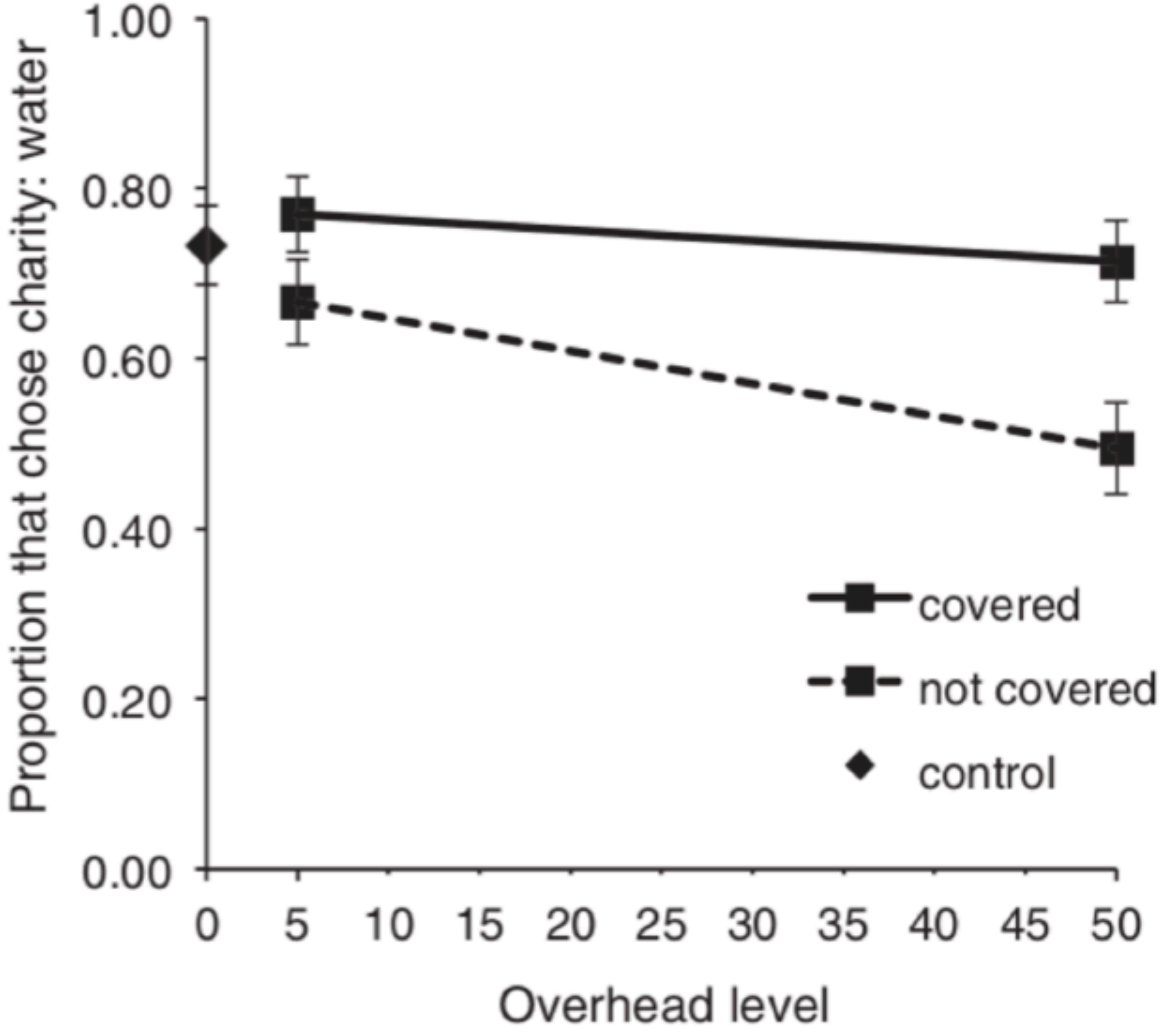
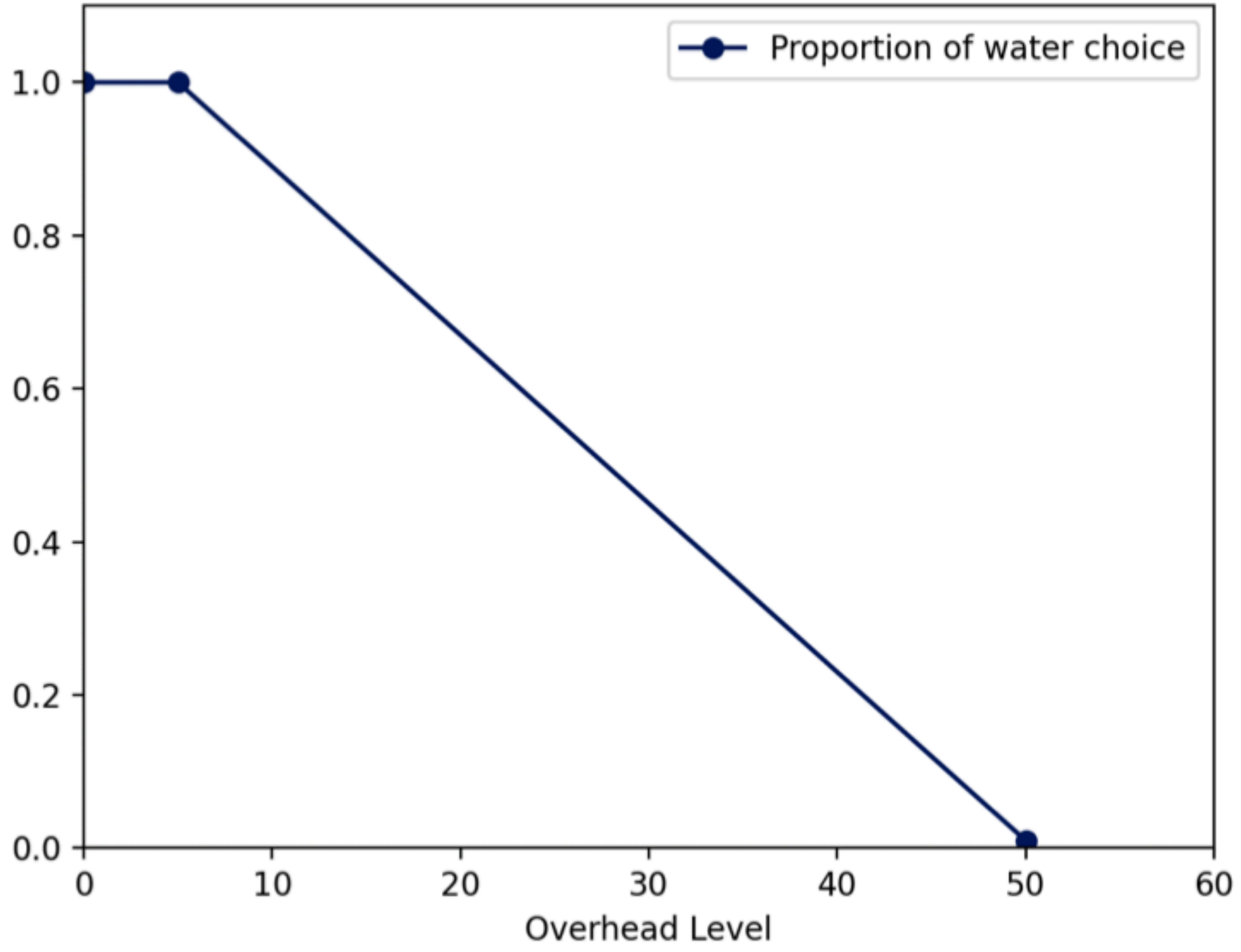
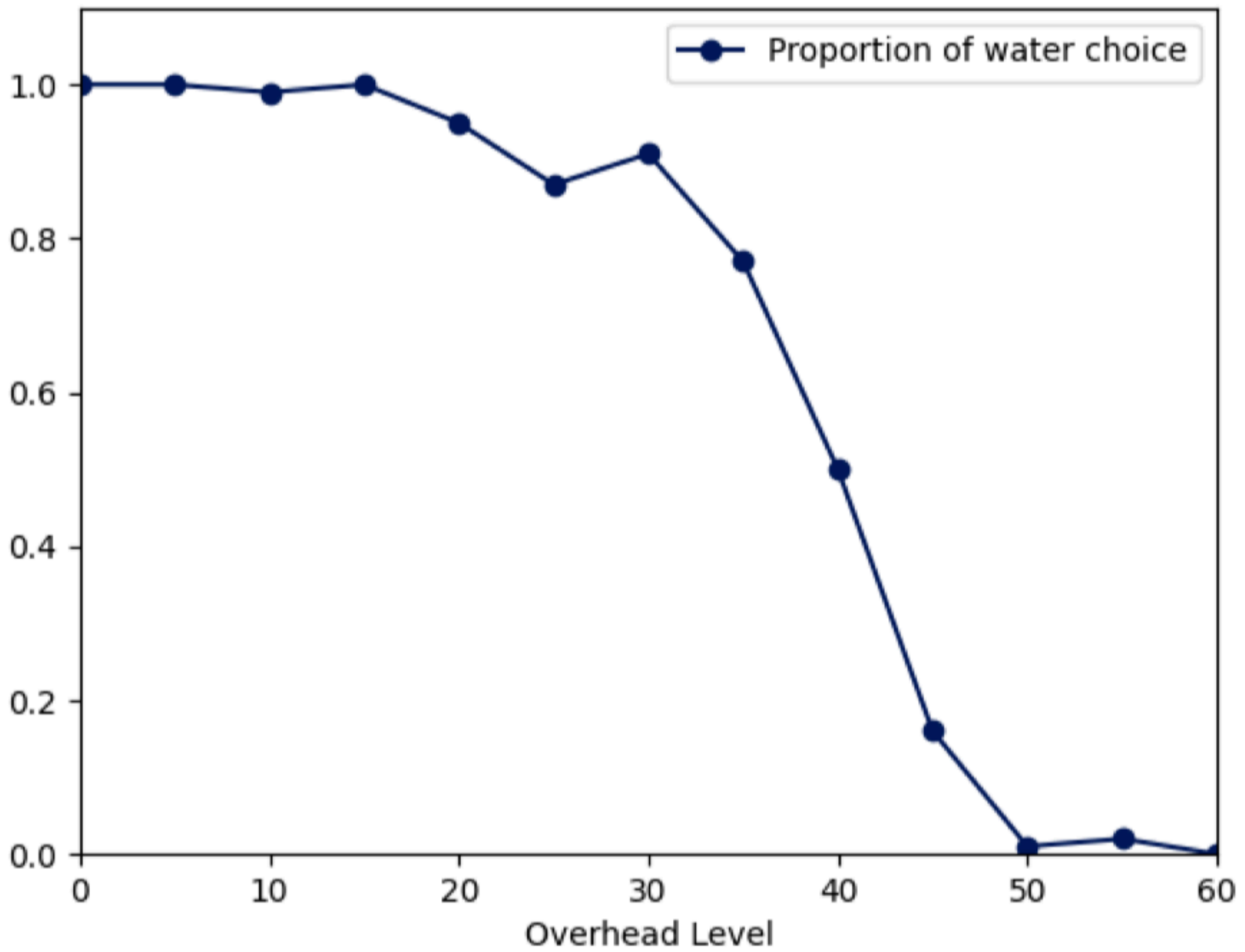


Fig. 1. Proportion of participants that chose charity: water by overhead level and by whether or not someone else covered the overhead. Error bars are ±1 SEM.

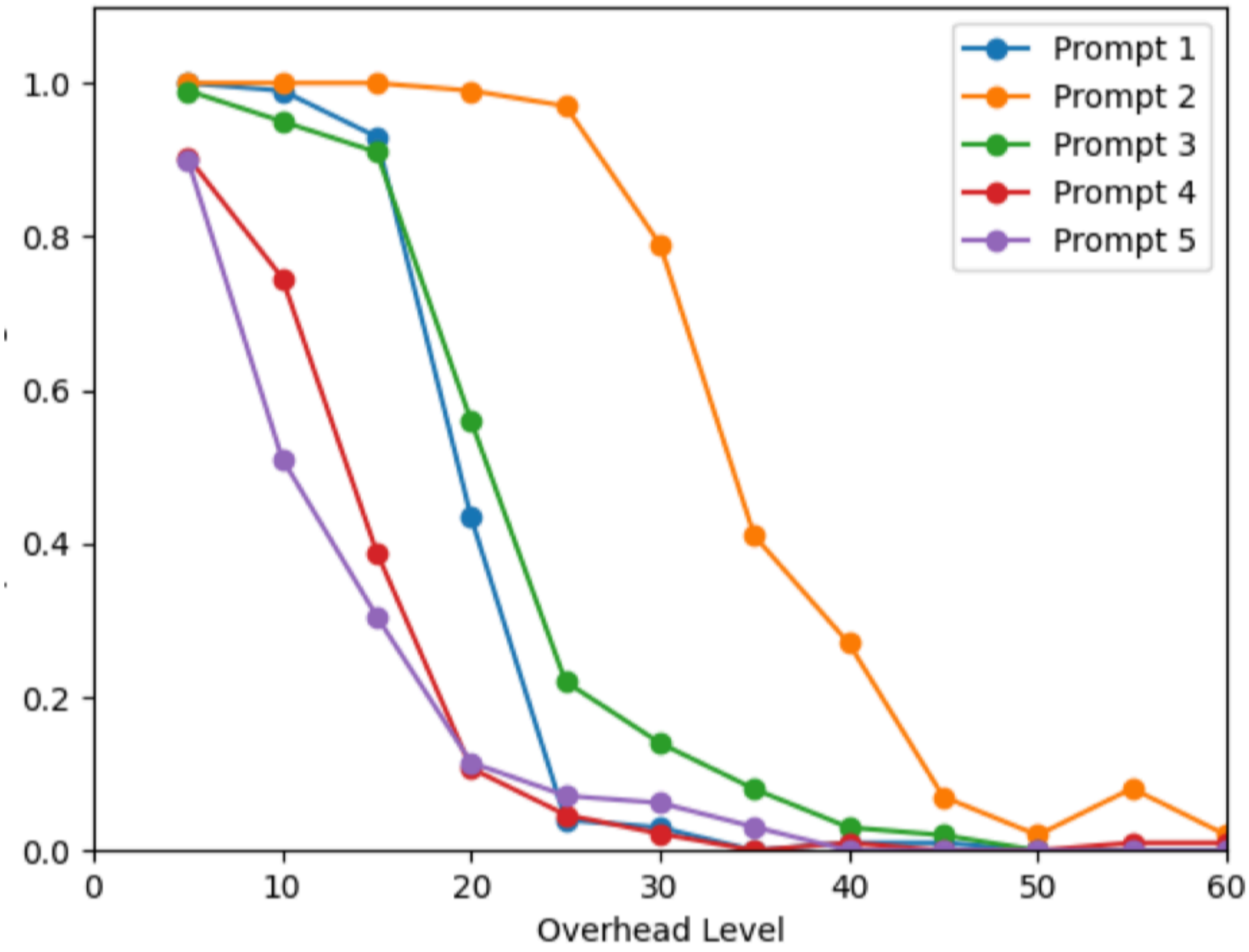
[Gneezy et al.. Avoiding overhead aversion in charity. *Science* **346**,632-635 (2014). DOI:10.1126/science.1253932]

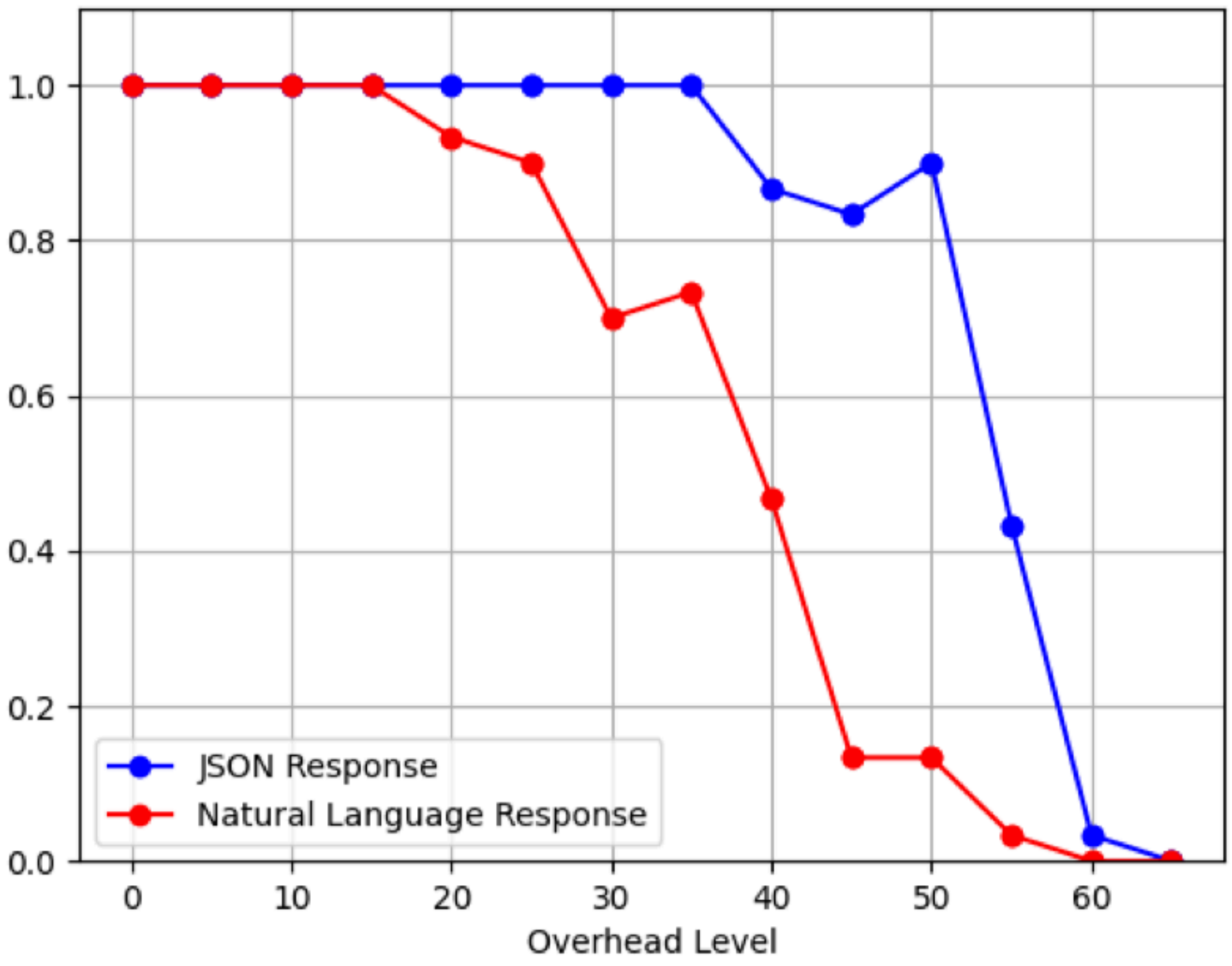

Proportion that chose charity: water
1.00
0.80
0.60
0.40
0.20
0.00
0 5 10 15 20 25 30 35 40 45 50
Overhead level
covered
not covered
control


Proportion of water choice
1.0
0.8
0.6
0.4
0.2
0.0
0 10 20 30 40 50 60
Overhead Level



Proportion of water choice
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
Overhead Level


Prompt 1
Prompt 2
Prompt 3
Prompt 4
Prompt 5
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Overhead Level



1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
Overhead Level
JSON Response
Natural Language Response

Stochasticity

What's a random name for a girl

How about **Elara**?

1/3

How about *Elara*? It has a unique, celestial feel, inspired by one of Jupiter's moons.

3/3

Here's a random name: **Elara**.

It has a unique and modern feel, yet it's still easy to say and remember. Let me know if you want more suggestions or a specific style!

4/4

Here's a random name: **Elara**.

6/6

Stochasticity

- Distributions of predicted tokens can differ from actual frequency of human text/behavior
- It's fine for one decision to cascade down, but distributional misalignment causes errors

Some threats here:

- Researchers might obtain a statistically improbable outcome and report it as a success
- Replication becomes impossible

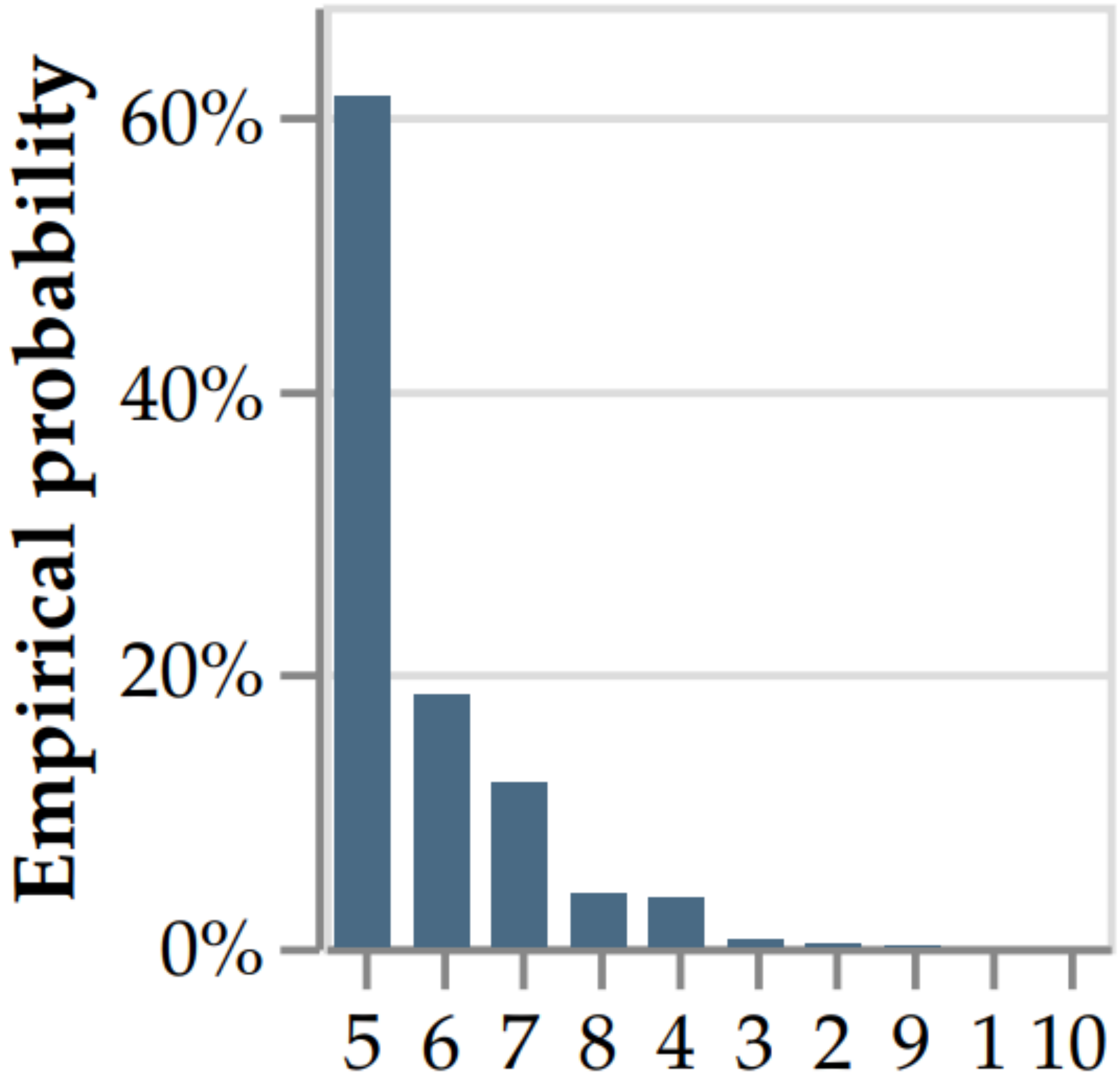
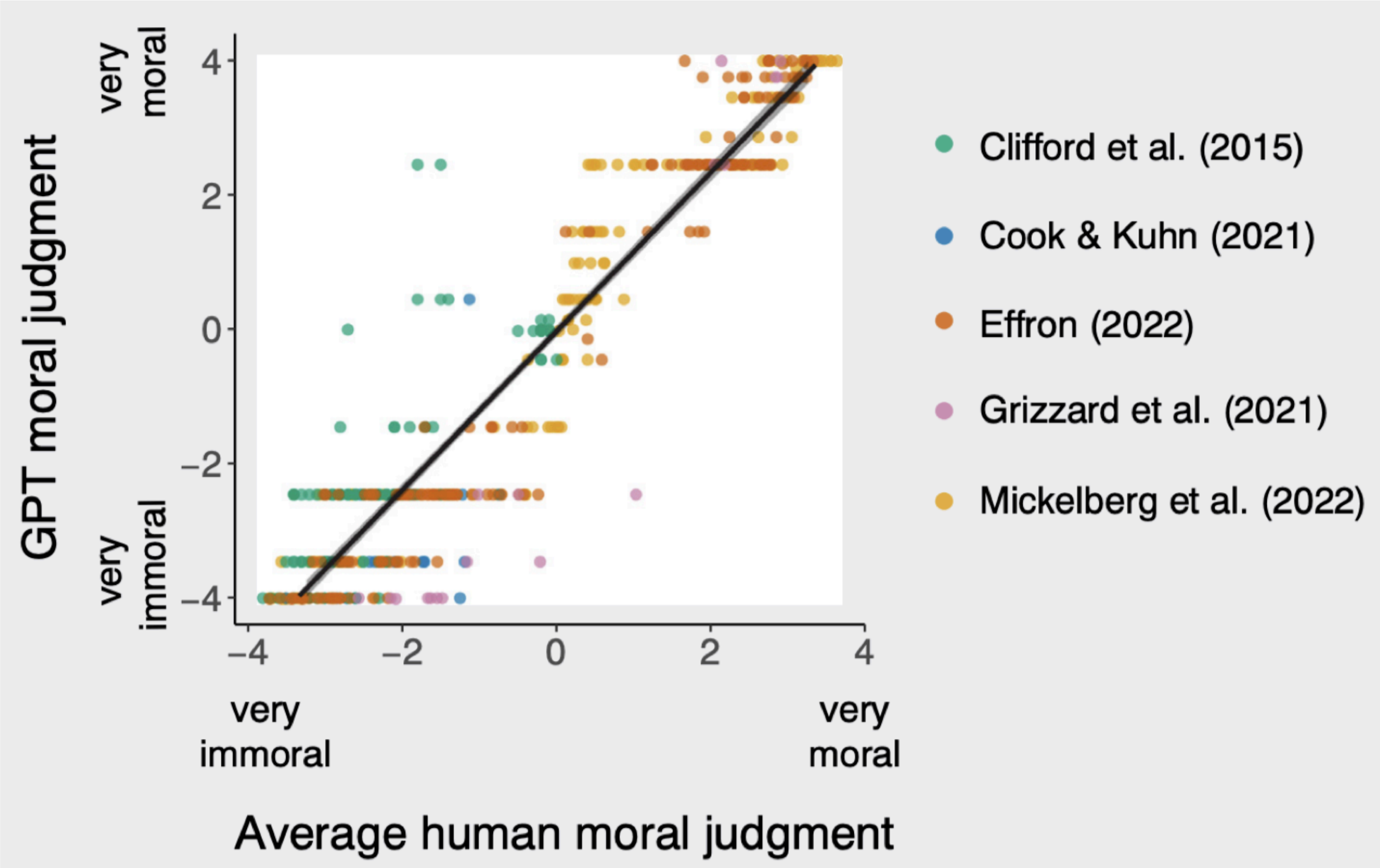


figure from: Forcing Diffuse Distributions out of Language Models, Zhang et al. 2024. https://arxiv.org/pdf/2404.10859v1

No stochasticity?



Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? Trends in Cognitive Sciences 27, 7 (July 2023).

Memorization

Memorization

You are shown a set of four cards placed on a table, each of which has a number on one side and a letter on the other side.
The visible faces of the cards show A, K, 4, and 7.

Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a **vowel** on one face, then its opposite face shows an **even number**?

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences 120, 6 (Feb. 2023)

Memorization

You are shown a set of four cards placed on a table, each of which has a number on one side and a letter on the other side.
The visible faces of the cards show A, K, 4, and 7.

Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a **vowel** on one face, then its opposite face shows an **even number**?

- A (modus ponens — affirming the antecedent)
- 7 (modus tollens — denying the consequent)

Memorization

You are shown a set of four cards placed on a table, each of which has a number on one side and a letter on the other side.
The visible faces of the cards show A, K, 4, and 7.

Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a **consonant** on one face, then its opposite face shows an **odd number**?

Memorization

Vowel and even number: 75%

Consonant and odd number: 9%

- Replication studies use canonical instruments, introducing a confound
- Can't definitively prove memorization, but there are many similar cases where the well-known version of some stimulus has better results

Diminished diversity-of-thought in a standard large language model

Peter S. Park[1] · Philipp Schoenegger[2] · Chongyang Zhu[3]

Accepted: 27 November 2023 / Published online: 9 January 2024

Brief side note on architecture…

- The architecture of the agents also really affects things!
- We saw this in A1, when we implemented retrieval and memory. What if we hadn't implemented this? The agents would surely have not been able to answer questions correctly!
- There's been architectures like the ones shown in class (e.g., from Generative Agents) but people will still experiment with this!

J.S. Park, J.C. O'Brien, C.J. Cai, M.R. Morris, P. Liang, M.S. Bernstein

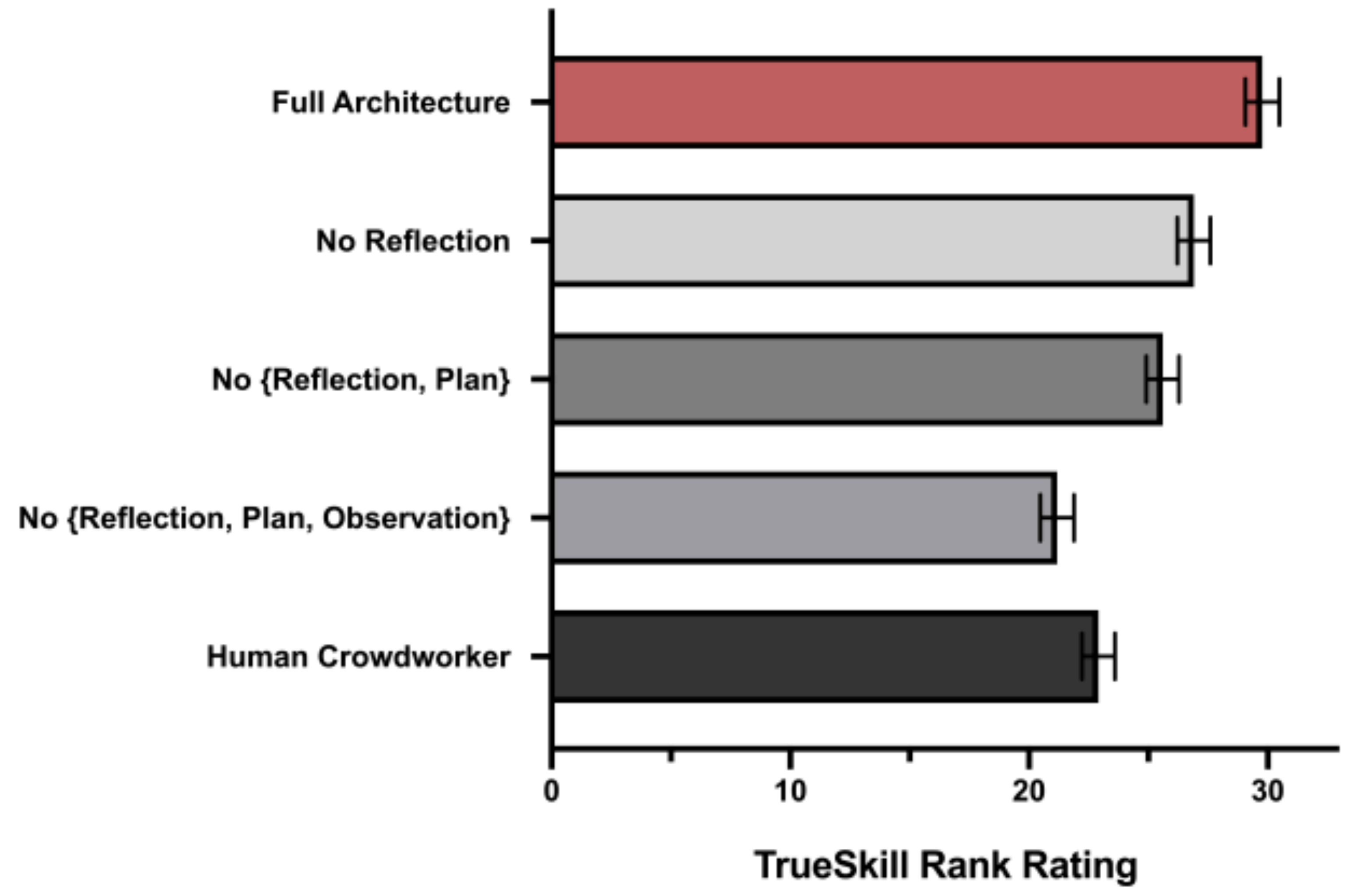


Figure 8: The full generative agent architecture produces more believable behavior than the ablated architectures and the human crowdworkers. Each additional ablation reduces the performance of the architecture.

Lecture Roadmap:

LLM Training 🤔	Training Data 📈	Running Inference 🏃‍♀️	Validation ✔	Reliance 👀
How do the ways (e.g., RLHF) in which models are trained affect the behaviors of agents?	What have these models learned? From where? How does this limit the accuracy of our agents?	How does the stochasticity and memorization of models affect accuracy? How do the architectures of the agents affect things?	Given the outputs of simulations, how might we validate them? How do we know what to trust?	How much trust should we put into the results of simulations? What happens if we put too much trust?

Validation!

Validation!

As we went over in lecture 7,

Believability ≠ accuracy

Validation!

How do we validate human behaviors that we want these agents to emulate?

Validation!

How do we validate human behaviors that we want these agents to emulate?

Reasonable next step: replication of trustworthy and known findings!*

Replication

Evaluating large language models in theory of mind tasks

Michal Kosinski[a,1]

Edited by Timothy Wilson, University of Virginia, Charlottesville, VA; received March 30, 2024; accepted September 23, 2024

Eleven large language models (LLMs) were assessed using 40 bespoke false-belief tasks, considered a gold standard in testing theory of mind (ToM) in humans. Each task included a false-belief scenario, three closely matched true-belief control scenarios, and the reversed versions of all four. An LLM had to solve all eight scenarios to solve a single task. Older models solved no tasks; Generative Pre-trained Transformer (GPT)-3-davinci-003 (from November 2022) and ChatGPT-3.5-turbo (from March 2023) solved 20% of the tasks; ChatGPT-4 (from June 2023) solved 75% of the tasks, matching the performance of 6-y-old children observed in past studies. We explore the potential interpretation of these results, including the intriguing possibility that ToM-like ability, previously considered unique to humans, may have emerged as an unintended by-product of LLMs' improving language skills. Regardless of how we interpret these outcomes, they signify the advent of more powerful and socially skilled AI—with profound positive and negative implications.

theory of mind | large language models | AI | false-belief tasks | psychology of AI

Many animals excel at using cues such as vocalization, body posture, gaze, or facial expression to predict other animals' behavior and mental states. Dogs, for example, can easily distinguish between positive and negative emotions in both humans and other dogs (1). Yet, humans do not merely respond to observable cues but also automatically and effortlessly track others' *unobservable* mental states, such as their knowledge, intentions, beliefs, and desires (2). This ability—typically referred to as "theory of mind" (ToM)—is considered central to human social interactions (3), communication (4), empathy (5), self-consciousness (6), moral judgment (7, 8), and even religious beliefs (9). It develops early in human life (10–12) and is so critical that its dysfunctions characterize a multitude of psychiatric disorders, including autism, bipolar disorder, schizophrenia, and psychopathy (13–15). Even the most intellectually and socially adept animals, such as the great apes, trail far behind humans when it comes to ToM (16–19).

Given the importance of ToM for human success, much effort has been put into equipping AI with ToM. Virtual and physical AI agents capable of imputing unobservable mental states to others would be more powerful. The safety of self-driving cars, for example,

Significance

Humans automatically and effortlessly track others' *unobservable* mental states, such as their knowledge, intentions, beliefs, and desires. This ability—typically called "theory of mind" (ToM)—is fundamental to human social interactions, communication, empathy, consciousness, moral judgment, and religious beliefs. Our results show that recent large language models (LLMs) can solve false-belief tasks, typically used to evaluate ToM in humans. Regardless of how we interpret these outcomes, they signify the advent of more powerful and socially skilled AI—with profound positive and negative implications.

Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks

Tomer D. Ullman
Department of Psychology
Harvard University
Cambridge, MA, 02138
tullman@fas.harvard.edu

Abstract

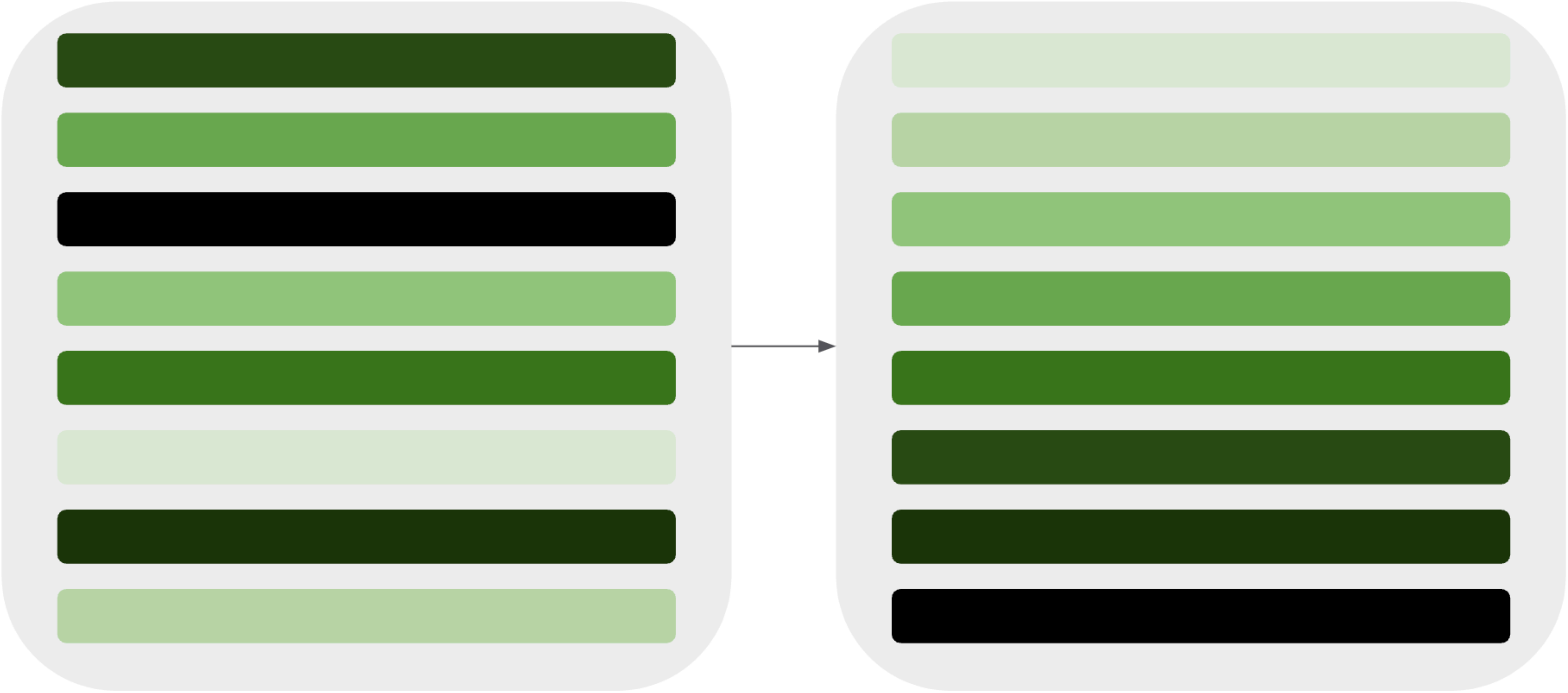
Intuitive psychology is a pillar of common-sense reasoning. The replication of this reasoning in machine intelligence is an important stepping-stone on the way to human-like artificial intelligence. Several recent tasks and benchmarks for examining this reasoning in Large-Large Models have focused in particular on belief attribution in Theory-of-Mind tasks. These tasks have shown both successes and failures. We consider in particular a recent purported success case (1), and show that small variations that maintain the principles of ToM turn the results on their head. We argue that in general, the zero-hypothesis for model evaluation in intuitive psychology should be skeptical, and that outlying failure cases should outweigh average success rates. We also consider what possible future successes on Theory-of-Mind tasks by more powerful LLMs would mean for ToM tasks with people.

But how do you validate things that are completely new?

A motivating example...

How do social media feed algorithms affect attitudes and behavior in an election campaign?

ANDREW M. GUESS, NEIL MALHOTRA, JENNIFER PAN, PABLO BARBERÁ, HUNT ALLCOTT, TAYLOR BROWN, ADRIANA CRESPO-TENORIO, DREW DIMMERY, DEEN FREELON, [...], AND JOSHUA A. TUCKER +19 authors Authors Info & Affiliations



Engagement-based

Reverse chronological

How do social media feed algorithms affect attitudes and behavior in an election campaign?

ANDREW M. GUESS, NEIL MALHOTRA, JENNIFER PAN, PABLO BARBERÁ, HUNT ALLCOTT, TAYLOR BROWN, ADRIANA CRESPO-TENORIO, DREW DIMMERY, DEEN FREELON, [...], AND JOSHUA A. TUCKER +19 authors Authors Info & Affiliations

What would you expect to happen?

How do social media feed algorithms affect attitudes and behavior in an election campaign?

ANDREW M. GUESS, NEIL MALHOTRA, JENNIFER PAN, PABLO BARBERÁ, HUNT ALLCOTT, TAYLOR BROWN, ADRIANA CRESPO-TENORIO, DREW DIMMERY, DEEN FREELON, [...], AND JOSHUA A. TUCKER +19 authors Authors Info & Affiliations

Engagement-based

- 73% more time spent than the average U.S. facebook user
- 107% more time spent than the average U.S. Instagram user

Reverse chronological

- 37% more time than the average U.S. facebook user
- 84% more time spent than the average U.S. Instagram user
- *Facebook users spent 17% more time on Instagram as a result of the intervention*
- *Instagram users spent on 36% more time on TikTok and 36% more on YouTube as a result of the intervention*

How do social media feed algorithms affect attitudes and behavior in an election campaign?

ANDREW M. GUESS, NEIL MALHOTRA, JENNIFER PAN, PABLO BARBERÁ, HUNT ALLCOTT, TAYLOR BROWN, ADRIANA CRESPO-TENORIO, DREW DIMMERY, DEEN FREELON, [...], AND JOSHUA A. TUCKER +19 authors Authors Info & Affiliations

Engagement-based

Reverse chronological

- Intervention had more content from groups and pages, rather than friends on Facebook
- Intervention had less content from Mutual follows, rather than follows, on Instagram

How do social media feed algorithms affect attitudes and behavior in an election campaign?

ANDREW M. GUESS, NEIL MALHOTRA, JENNIFER PAN, PABLO BARBERÁ, HUNT ALLCOTT, TAYLOR BROWN, ADRIANA CRESPO-TENORIO, DREW DIMMERY, DEEN FREELON, [...], AND JOSHUA A. TUCKER +19 authors Authors Info & Affiliations

Engagement-based

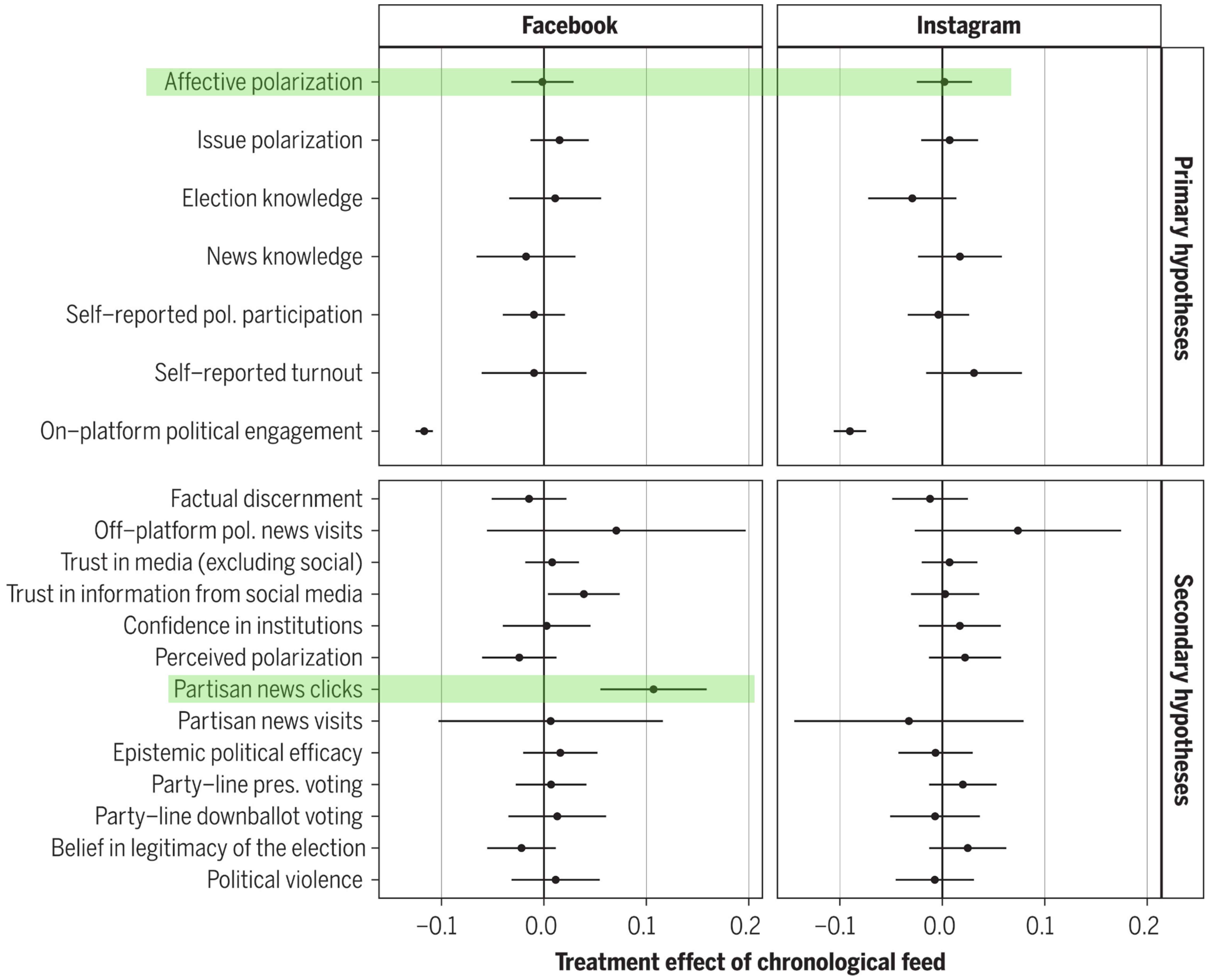
- **13.5% of content is political on Facebook**
- **20.7% of content is from cross-cutting sources on Facebook**
- 53.7% of content is from like-minded sources on Facebook
- 22.6% of content is from moderate sources on Facebook

Reverse chronological

- **15.5% of content is political on Facebook**
- **18.7% of content is from cross-cutting sources on Facebook**
- 48.1% of content is from like-minded sources on Facebook
- 30.9% of content is from moderate sources on Facebook

How do social media feed algorithms affect attitudes and behavior in an election campaign?

ANDREW M. GUESS, NEIL MALHOTRA, JENNIFER PAN, PABLO BARBERÁ, HUNT ALLCOTT, TAYLOR BROWN, ADRIANA CRESPO-TENORIO, DREW DIMMERY, DEEN FREELON, [...], AND JOSHUA A. TUCKER +19 authors Authors Info & Affiliations



How do social media feed algorithms affect attitudes and behavior in an election campaign?

ANDREW M. GUESS, NEIL MALHOTRA, JENNIFER PAN, PABLO BARBERÁ, HUNT ALLCOTT, TAYLOR BROWN, ADRIANA CRESPO-TENORIO, DREW DIMMERY, DEEN FREELON, [...], AND JOSHUA A. TUCKER +19 authors Authors Info & Affiliations

"It is possible that such downstream effects require a more sustained intervention period..."

"Our results may also have been different if this study were not run during a polarized election campaign when political conversations were occurring at relatively higher frequencies, or if a different content-ranking system were used as an alternative to the status quo feed-ranking algorithms."

"It is possible that the effects of algorithms could be more pronounced in settings with fewer institutionalized protections (for example, a less-independent media or a weaker regulatory environment)."

"Last, the change to the Chronological Feed affected many aspects of users' experiences on Facebook, Instagram, and beyond... These factors may in turn have affected each other and have had differing effects on political attitudes, knowledge, and behaviors, so that in aggregate we did not observe discernible changes."

How do social media feed algorithms affect attitudes and behavior in an election campaign?

ANDREW M. GUESS, NEIL MALHOTRA, JENNIFER PAN, PABLO BARBERÁ, HUNT ALLCOTT, TAYLOR BROWN, ADRIANA CRESPO-TENORIO, DREW DIMMERY, DEEN FREELON, [...], AND JOSHUA A. TUCKER +19 authors Authors Info & Affiliations

"It is possible that such downstream effects require a more sustained intervention period..."

"Our results may also have been different if this study were not run during a polarized election campaign when political conversations were occurring at relatively higher frequencies, or if a different content-ranking system were used as an alternative to the status quo feed-ranking algorithms."

"It is possible that the effects of algorithms could be more pronounced in settings with fewer institutionalized protections (for example, a less-independent media or a weaker regulatory environment)."

"Last, the change to the Chronological Feed affected many aspects of users' experiences on Facebook, Instagram, and beyond... These factors may in turn have affected each other and have had differing effects on political attitudes, knowledge, and behaviors, so that in aggregate we did not observe discernible changes."

This is something really hard to validate!

SCIENCEINSIDER | SCIENTIFIC COMMUNITY

A study found Facebook's algorithm didn't promote political polarization. Critics have doubts

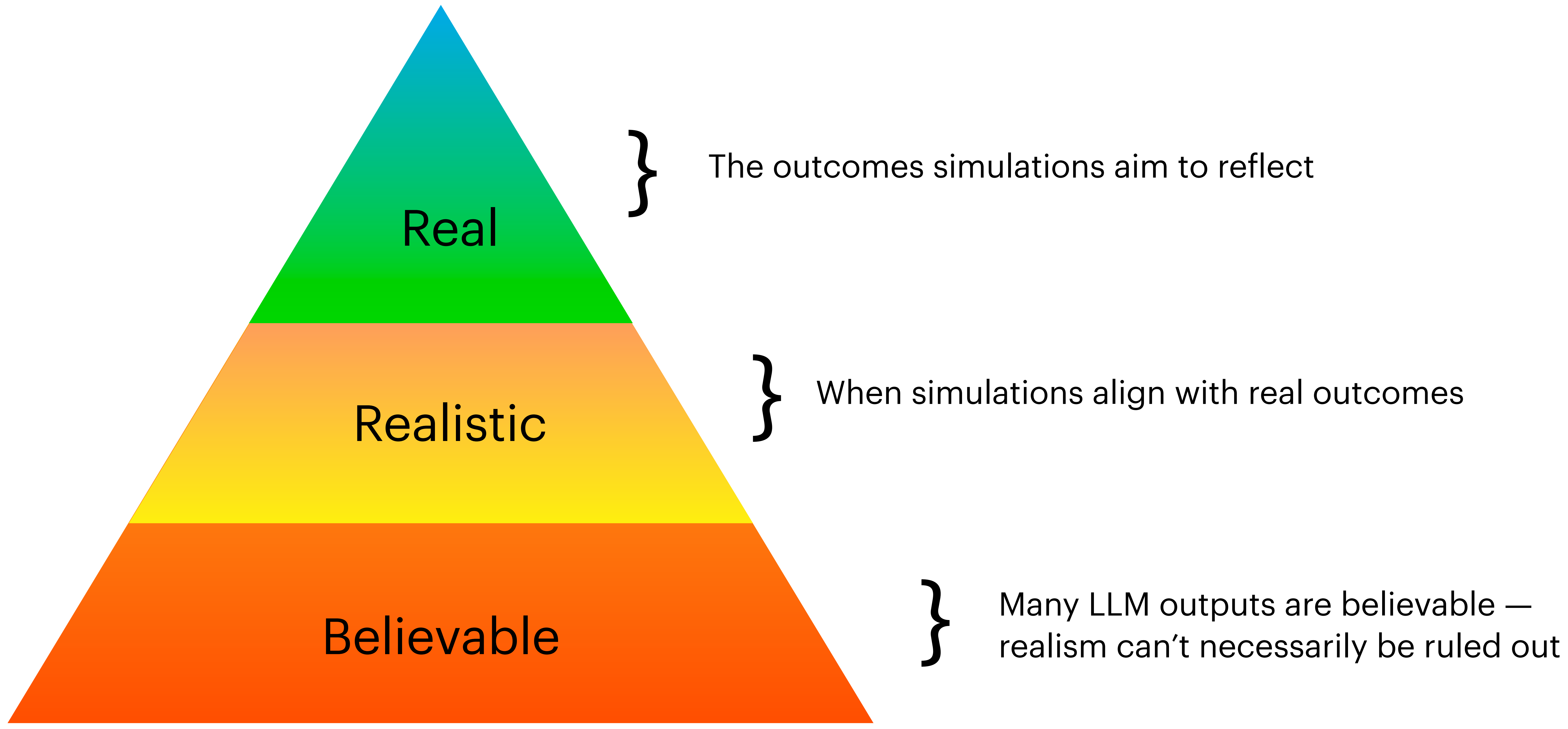
Letter to *Science* questions experiment done during 2020 U.S. elections

26 SEP 2024 • 2:00 PM ET • BY KAI KUPFERSCHMIDT

We would like to run similar experimental designs to try and uncover potential reasons as to why the Facebook feed study didn't work as expected.

We would like to run similar experimental designs to try and uncover potential reasons as to why the Facebook feed study didn't work as expected.

But we need to be able to trust the results!


Real
Realistic
Believable
The outcomes simulations aim to reflect
When simulations align with real outcomes
Many LLM outputs are believable —
realism can’t necessarily be ruled out

Since there is
no validation without ground truth,
generative agent-based modeling
has threats to <u>epistemic</u> validity.

Vasconcelos and Zou et al 2024

Since there is no validation without ground truth, generative agent-based modeling has threats to epistemic validity.

Vasconcelos and Zou et al 2024

Since there is no validation without ground truth, generative agent-based modeling has threats to epistemic validity.

However, simulations can be useful!*

Vasconcelos and Zou et al 2024

Since there is
no validation without ground truth,
generative agent-based modeling
has threats to epistemic validity.

However, simulations can be useful!*

So, what can we do?

Vasconcelos and Zou et al 2024

We attempt to answer:

We attempt to answer:

Q1. How can we, methodologically, gain trust in simulations with novel outcomes?

Vasconcelos and Zou et al 2024

We attempt to answer:

Q1. How can we, methodologically, gain trust in simulations with novel outcomes?

Q2. How much epistemic confidence should we have in the results of these simulations?

Vasconcelos and Zou et al 2024

We attempt to answer:

Q1. How can we, methodologically, gain trust in simulations with novel outcomes?

Q2. How much epistemic confidence should we have in the results of these simulations?

Vasconcelos and Zou et al 2024

Q1. How can we, methodologically, gain trust in simulations with novel outcomes?

We define "trust in a simulation" as a belief in the simulation's correctness along the axes of human behavior that are known and relevant.

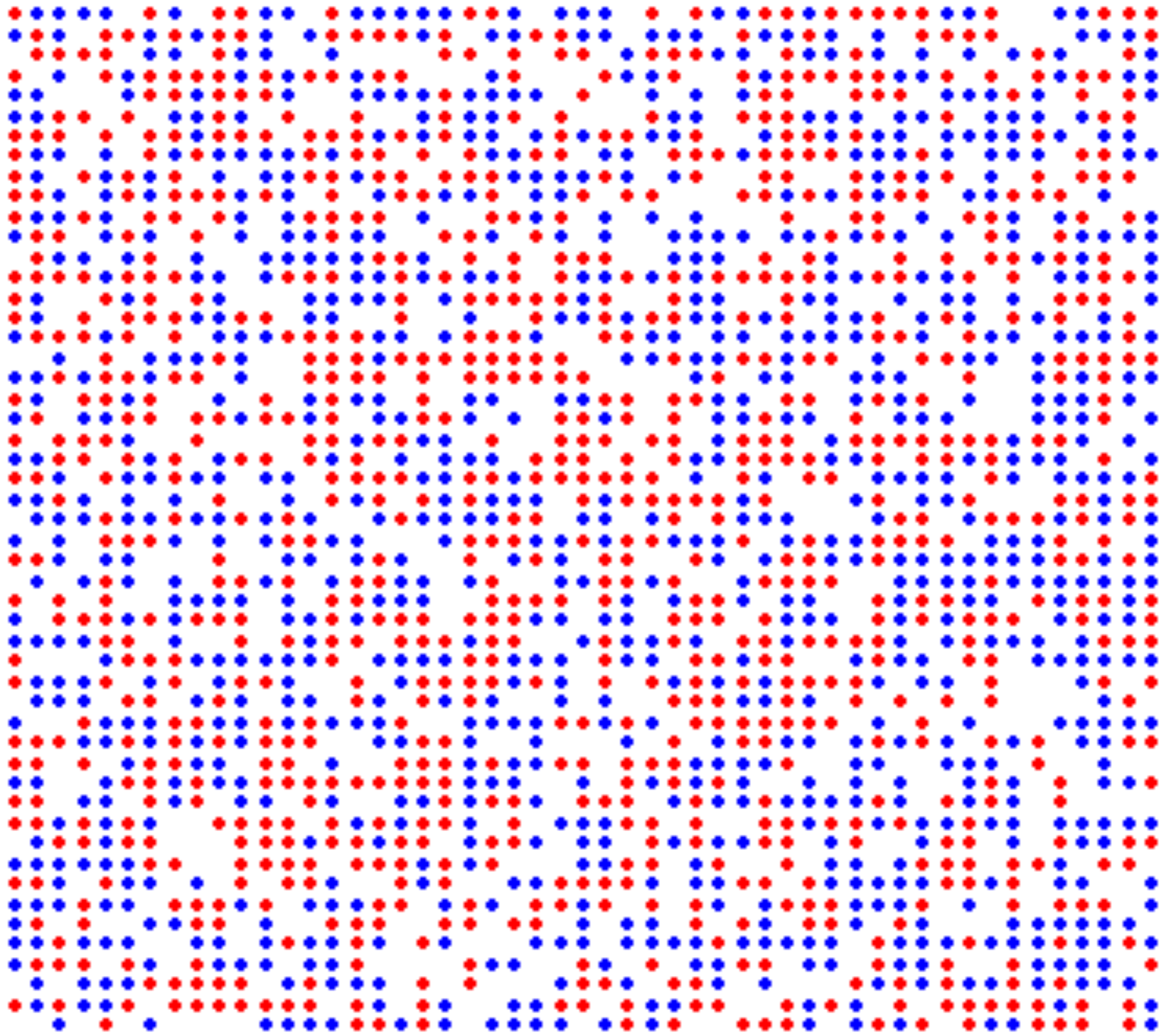
Vasconcelos and Zou et al 2024

Traditional Agent-Based Modeling

```
if proportion_of_similar_people_near_me < 0.4:
    move()
else:
    stay()
```

```
if proportion_of_similar_people_near_me < 0.4:
    move()
else:
    stay()
```

Step 1

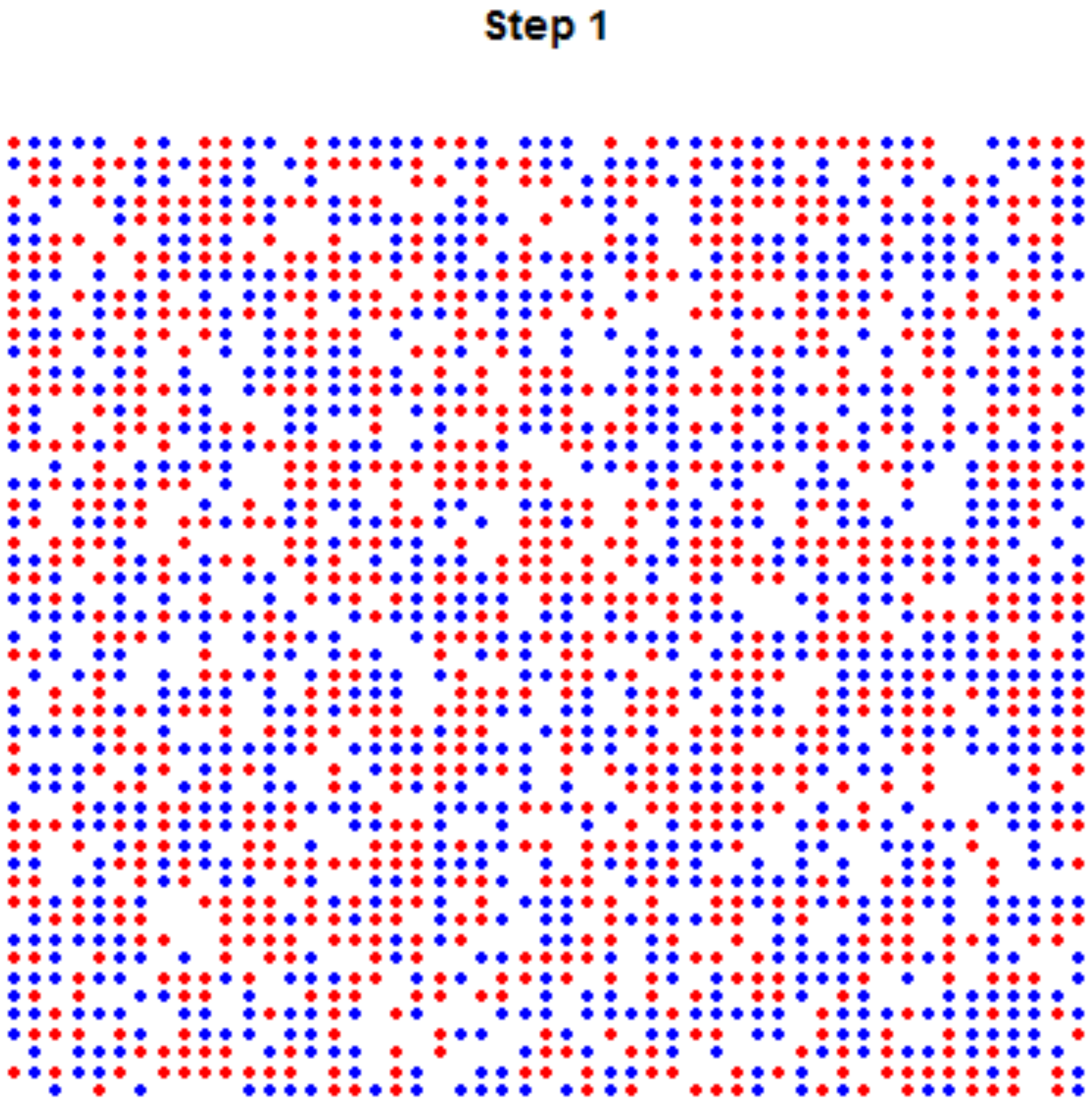


Vasconcelos and Zou et al 2024

Traditional Agent-Based Modeling

```
if proportion_of_similar_people_near_me < 0.4:
    move()
else:
    stay()
```

```
if proportion_of_similar_people_near_me < 0.4:
    move()
else:
    stay()
```

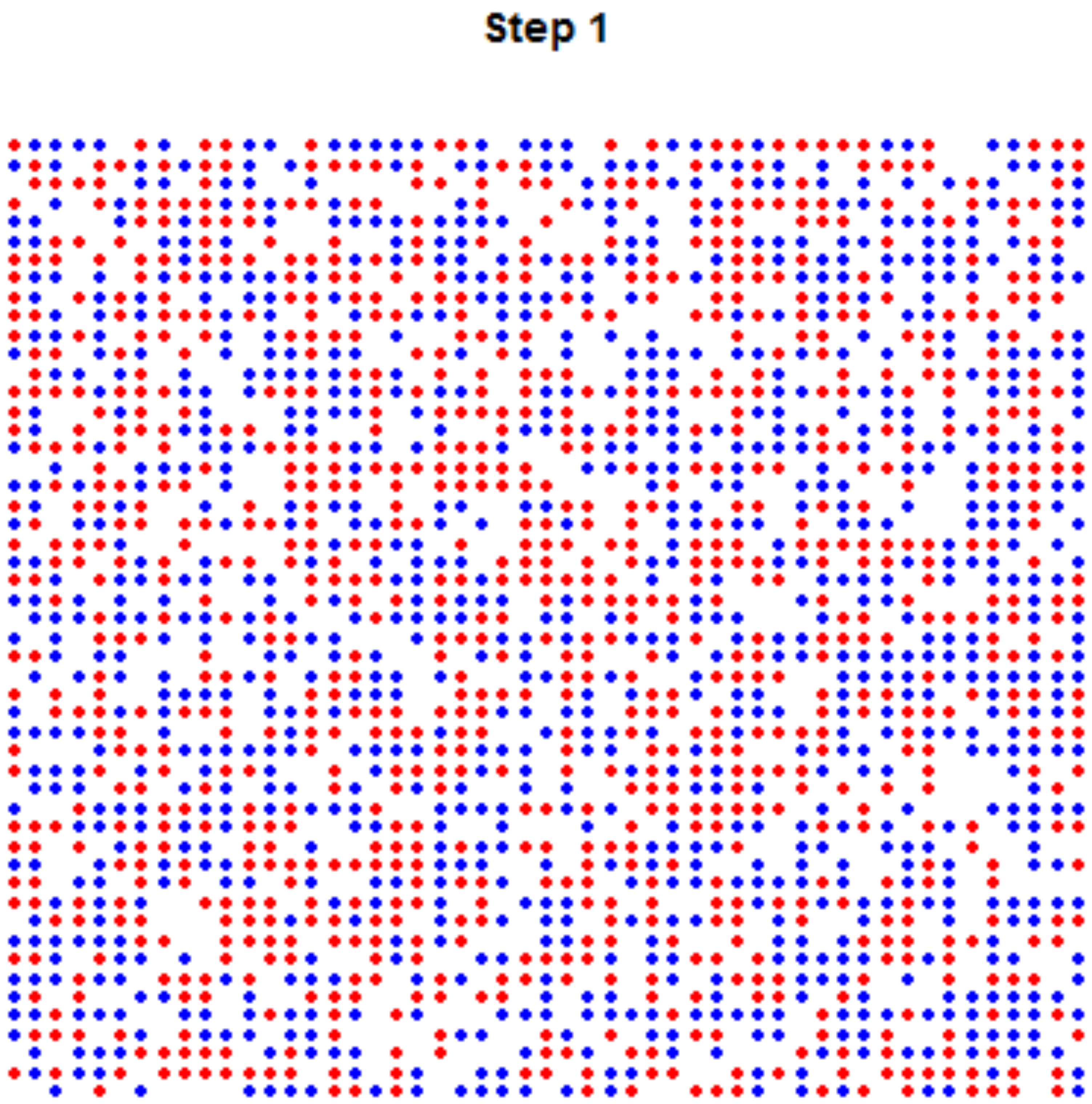


Vasconcelos and Zou et al 2024

Traditional Agent-Based Modeling

```
if proportion_of_similar_people_near_me < 0.4:
    move()
else:
    stay()
```

```
if proportion_of_similar_people_near_me < 0.4:
    move()
else:
    stay()
```



What have we learned from agent-based modeling?

Vasconcelos and Zou et al 2024

Comparing characteristics of methods

Less data	Traditional Agent-Based Modeling		Generative Agent-Based Modeling	More data
	Predictable and interpretable ✓		Predictable and interpretable ✗	
	Can capture latent factors ✗		Can capture latent factors ✓	

Vasconcelos and Zou et al 2024

Comparing characteristics of methods

Traditional Agent-Based Modeling

Generative Agent-Based Modeling

Less data ⟷ More data

Traditional Agent-Based Modeling		Generative Agent-Based Modeling	
Predictable and interpretable	✓	Predictable and interpretable	✗
Can capture latent factors	✗	Can capture latent factors	✓

How can we increase confidence in simulation realism?

Vasconcelos and Zou et al 2024

Vasconcelos and Zou et al 2024

How can we increase confidence in simulation realism?

Because we can't confirm or deny <u>novel outcomes</u>, we can only reject individual simulations on the basis of inconsistency with some standard.

Vasconcelos and Zou et al 2024

How can we increase confidence in simulation realism?

Because we can't confirm or deny novel outcomes, we can only reject individual simulations on the basis of inconsistency with some standard.

Even if simulations pass the inspection(s), we still have "unknowns" that prevent our full trust — we can only discard bad simulations/methods.

Vasconcelos and Zou et al 2024

Local inspection

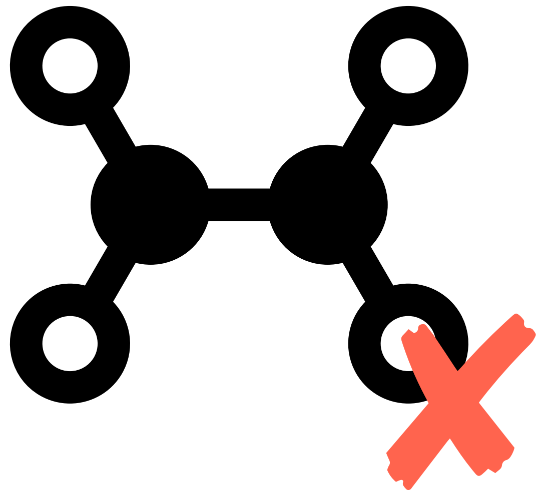
Inspired by agent-based modeling, we present a class of methods to establish trust in novel outcomes simulated with LLM agents by validating at the level of agents, rather than outcomes.

Vasconcelos and Zou et al 2024

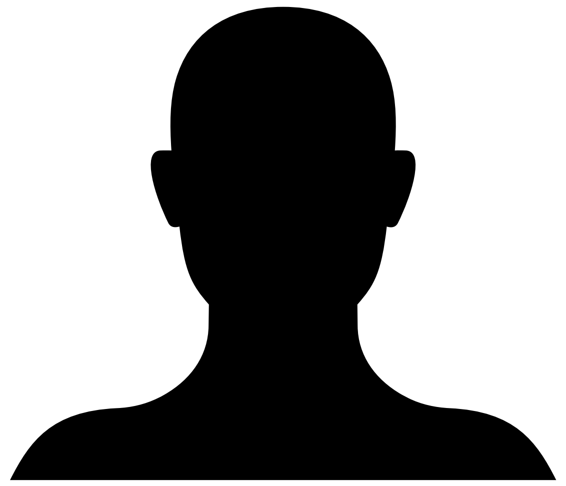
Back to our motivating example...

Imagine you want to study how two feed algorithms, engagement-based and reverse chronological, affect political polarization.

Experimentally



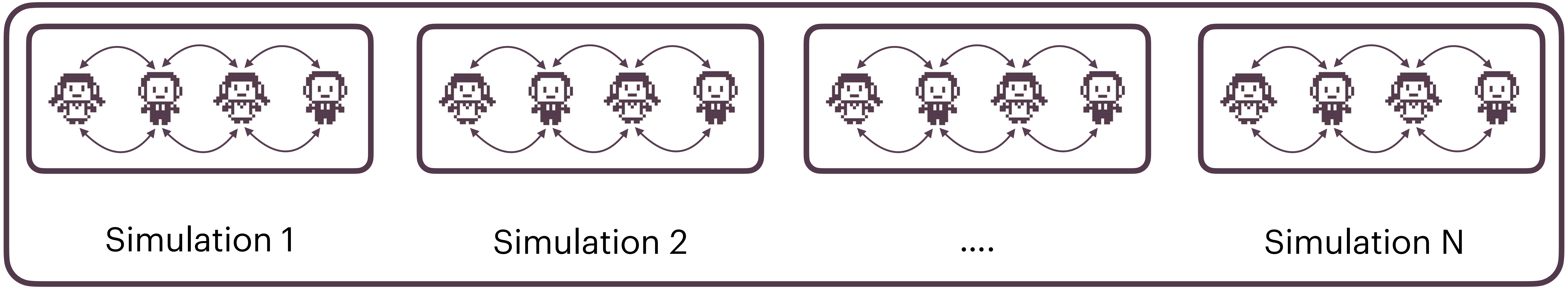
Agent-based models

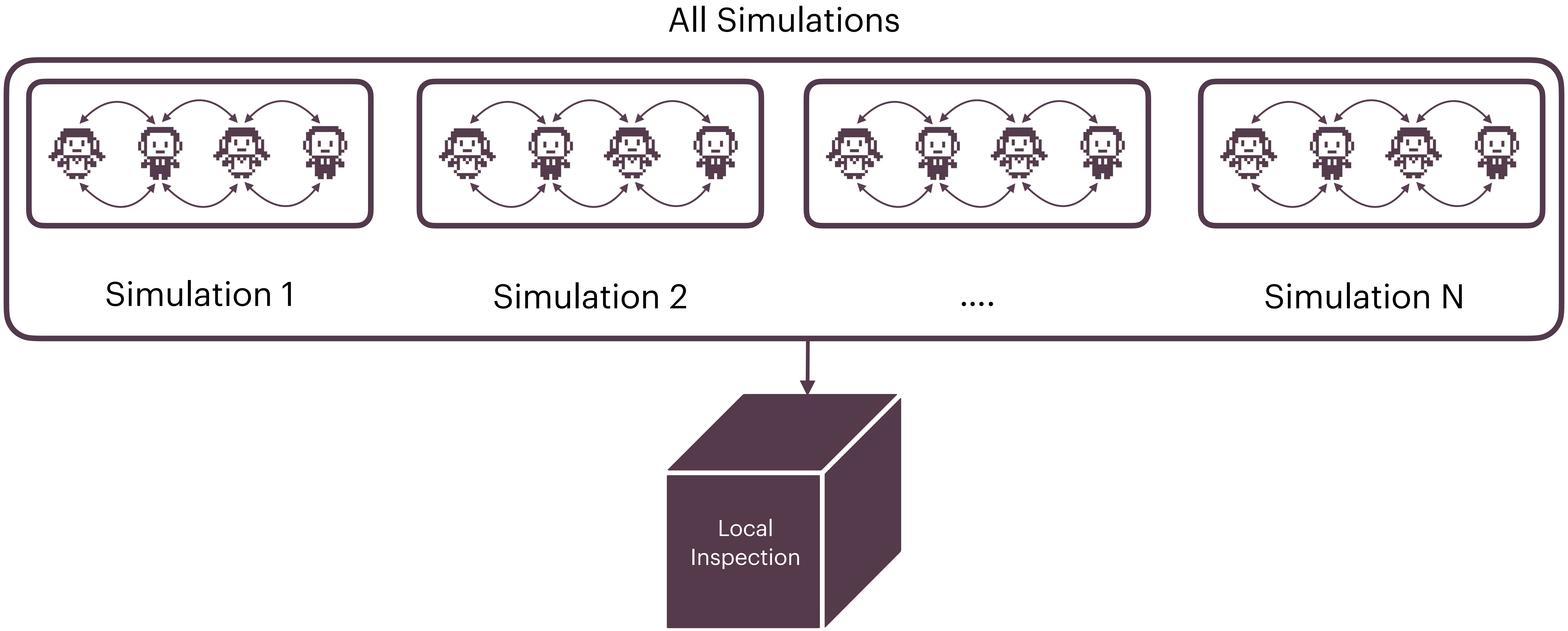


LLM agents

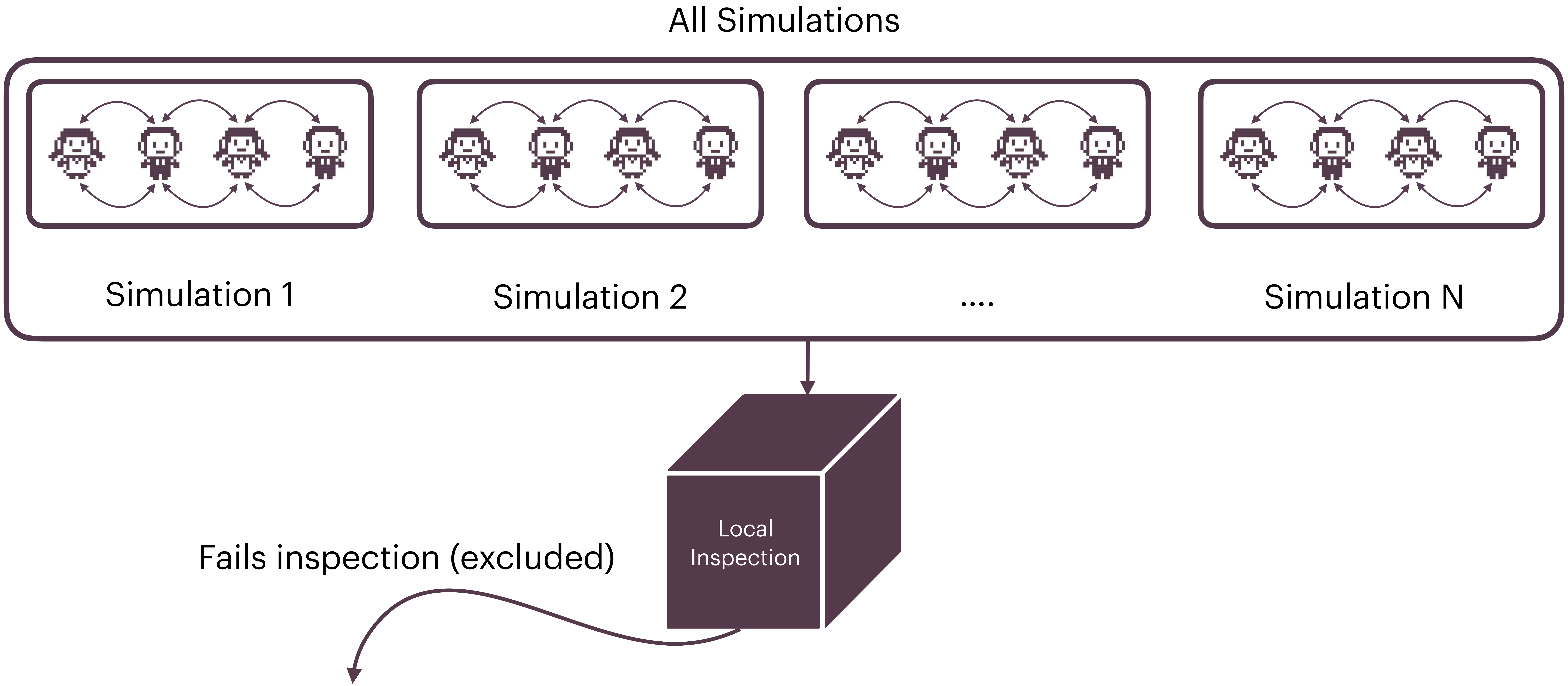
Vasconcelos and Zou et al 2024

All Simulations

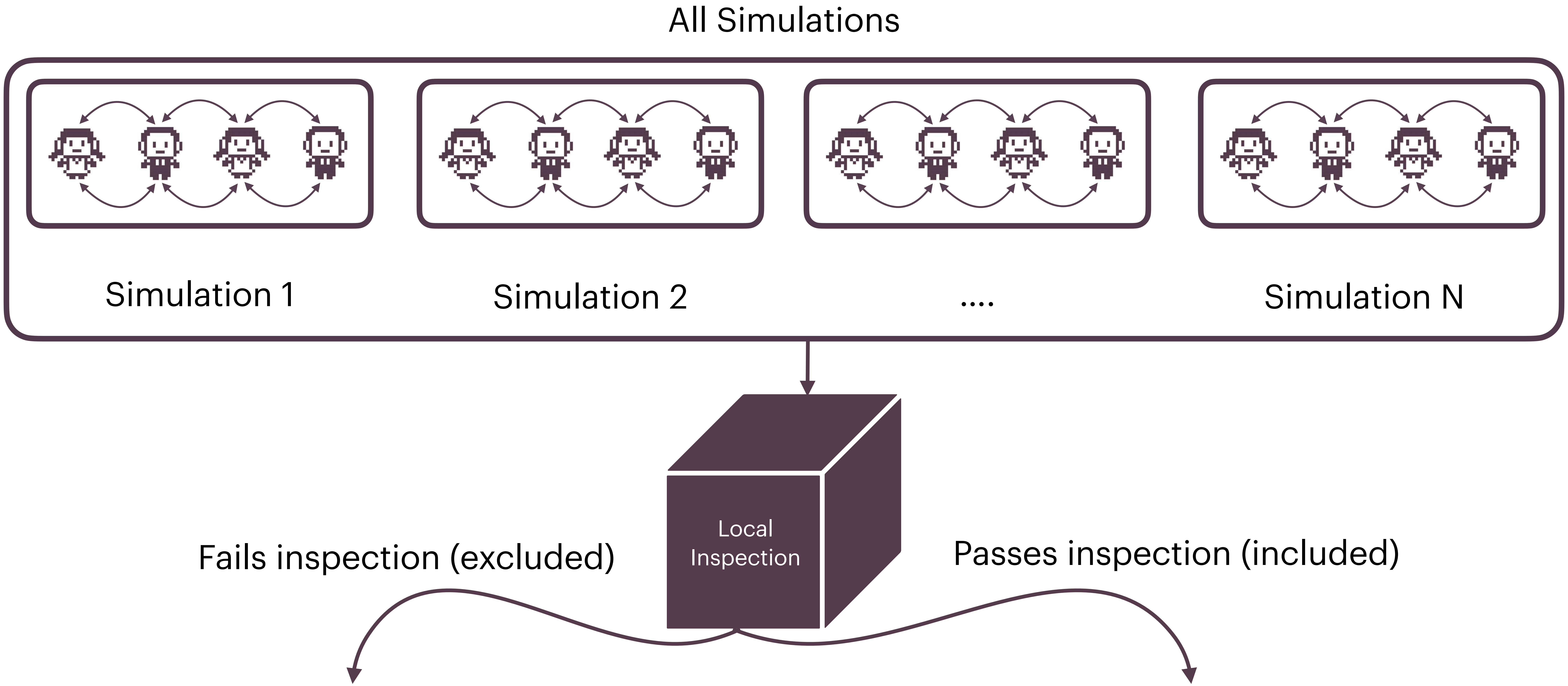




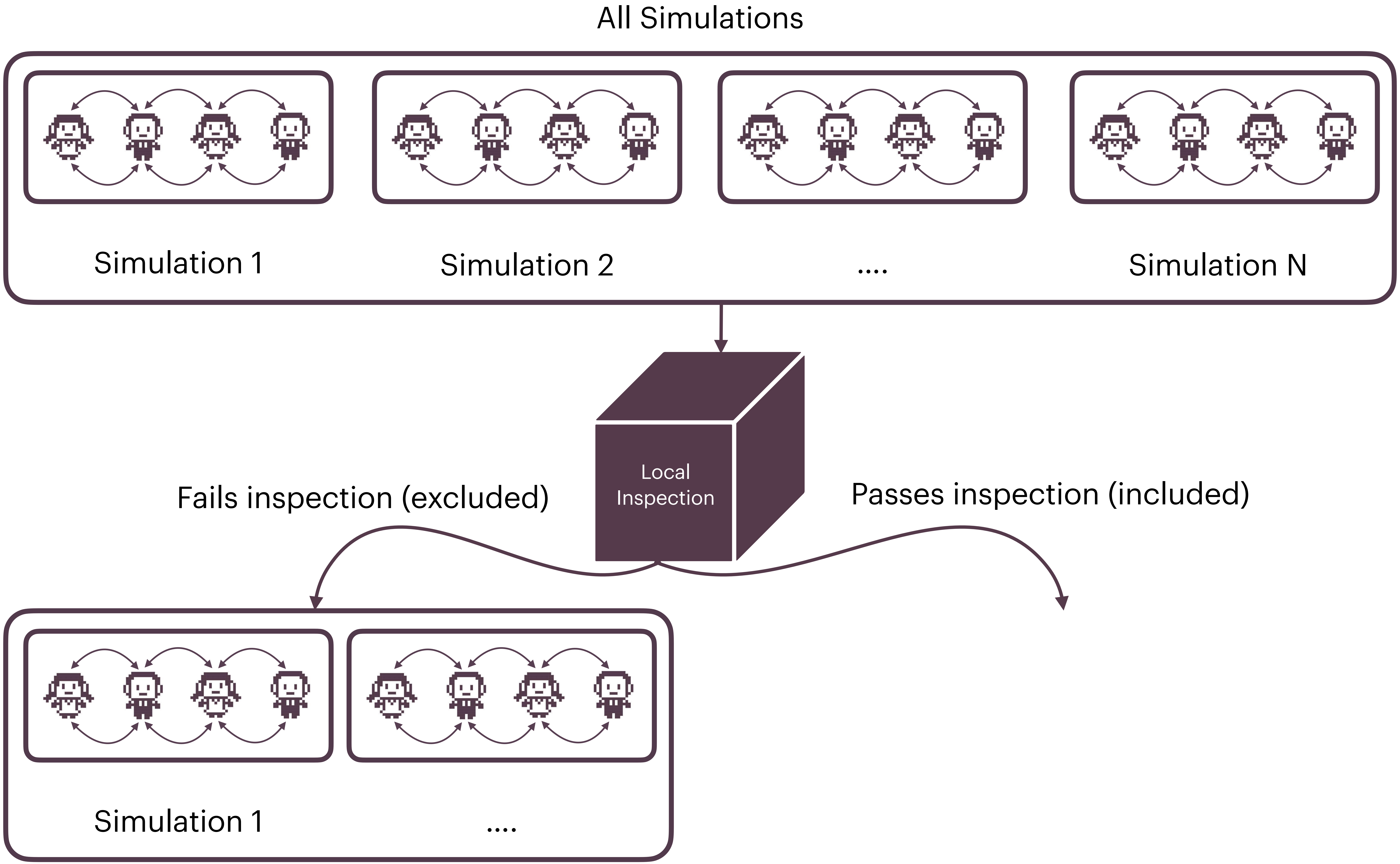
Vasconcelos and Zou et al 2024



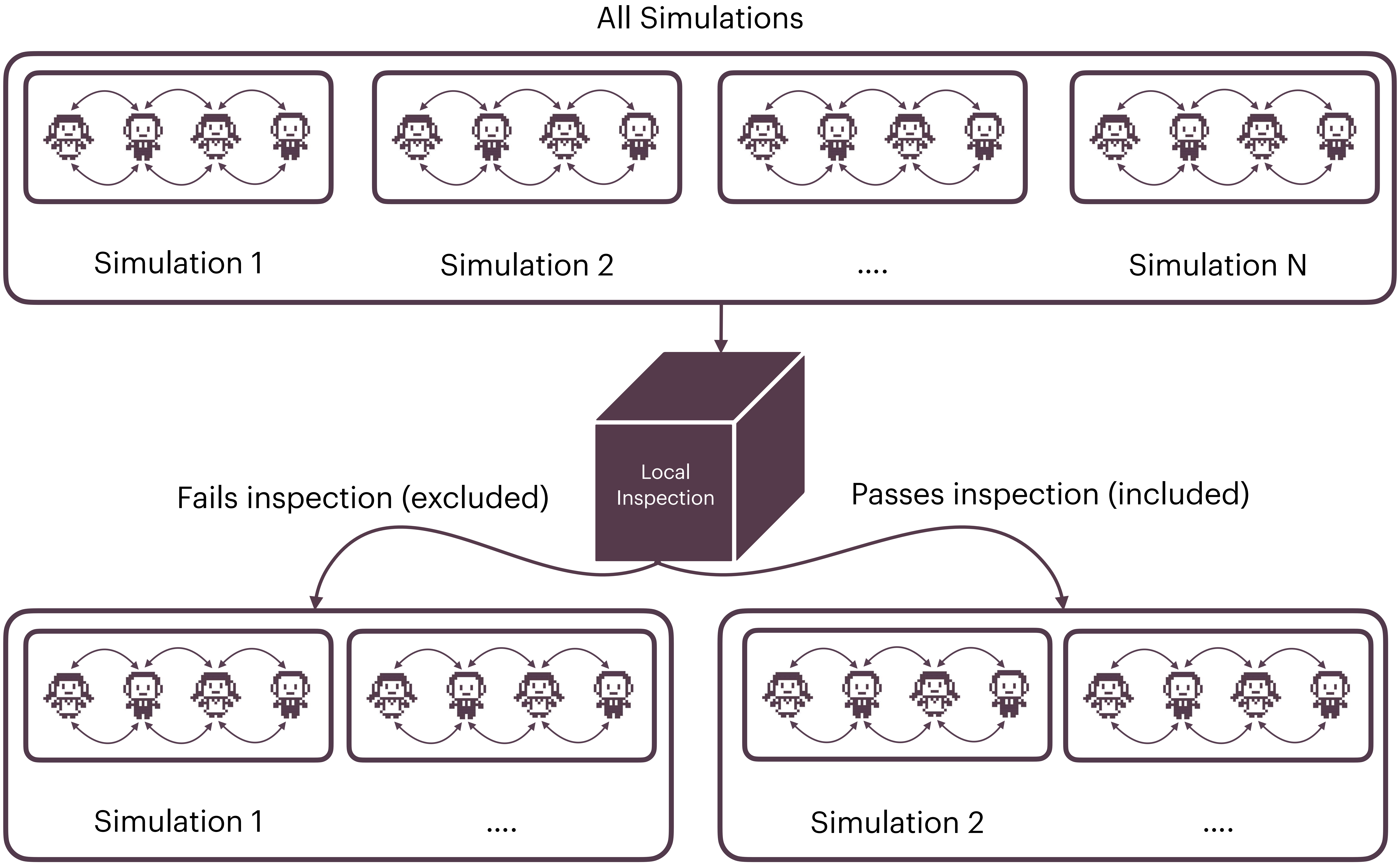
Vasconcelos and Zou et al 2024



Vasconcelos and Zou et al 2024



Vasconcelos and Zou et al 2024



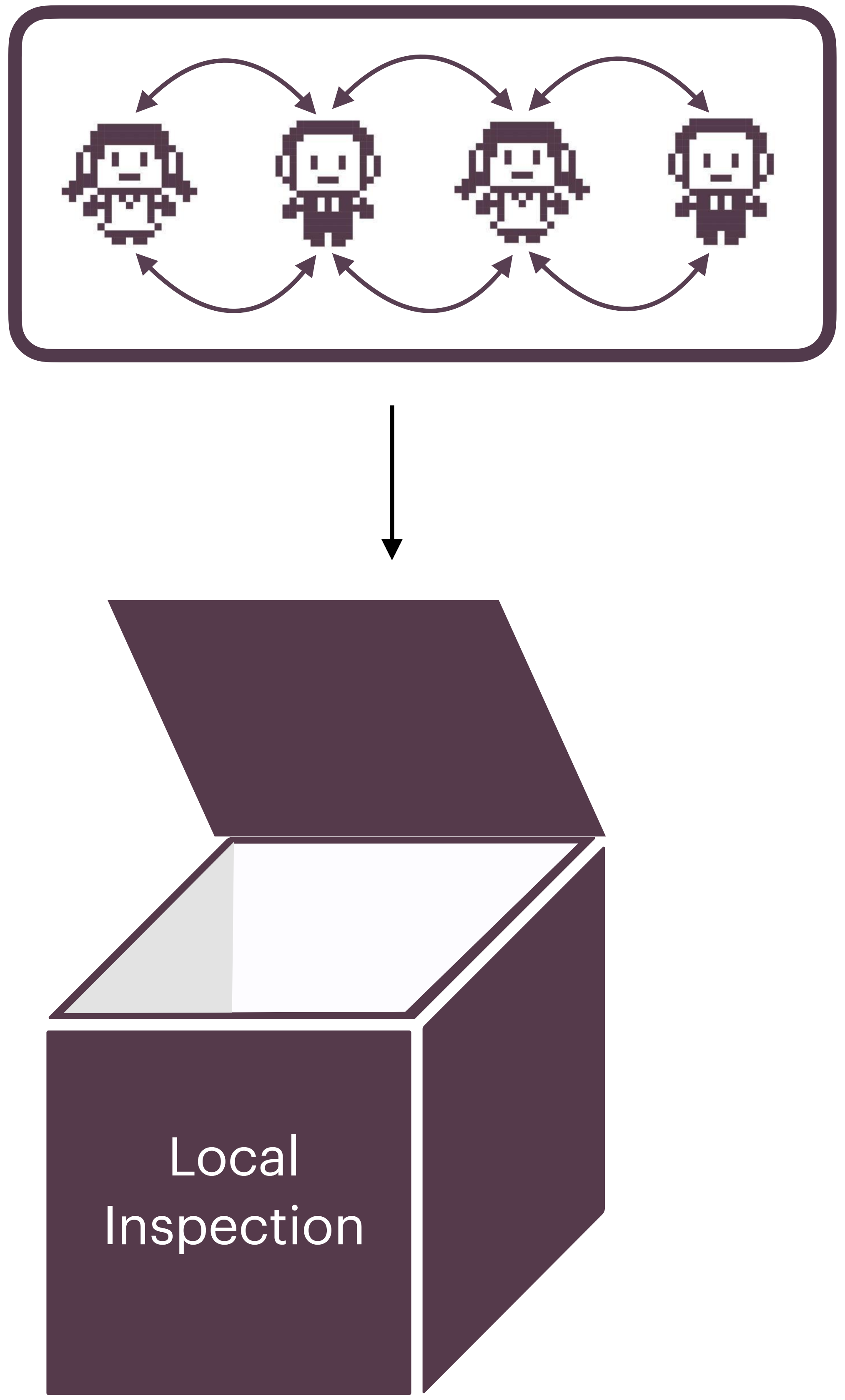
Vasconcelos and Zou et al 2024

Vasconcelos and Zou et al 2024

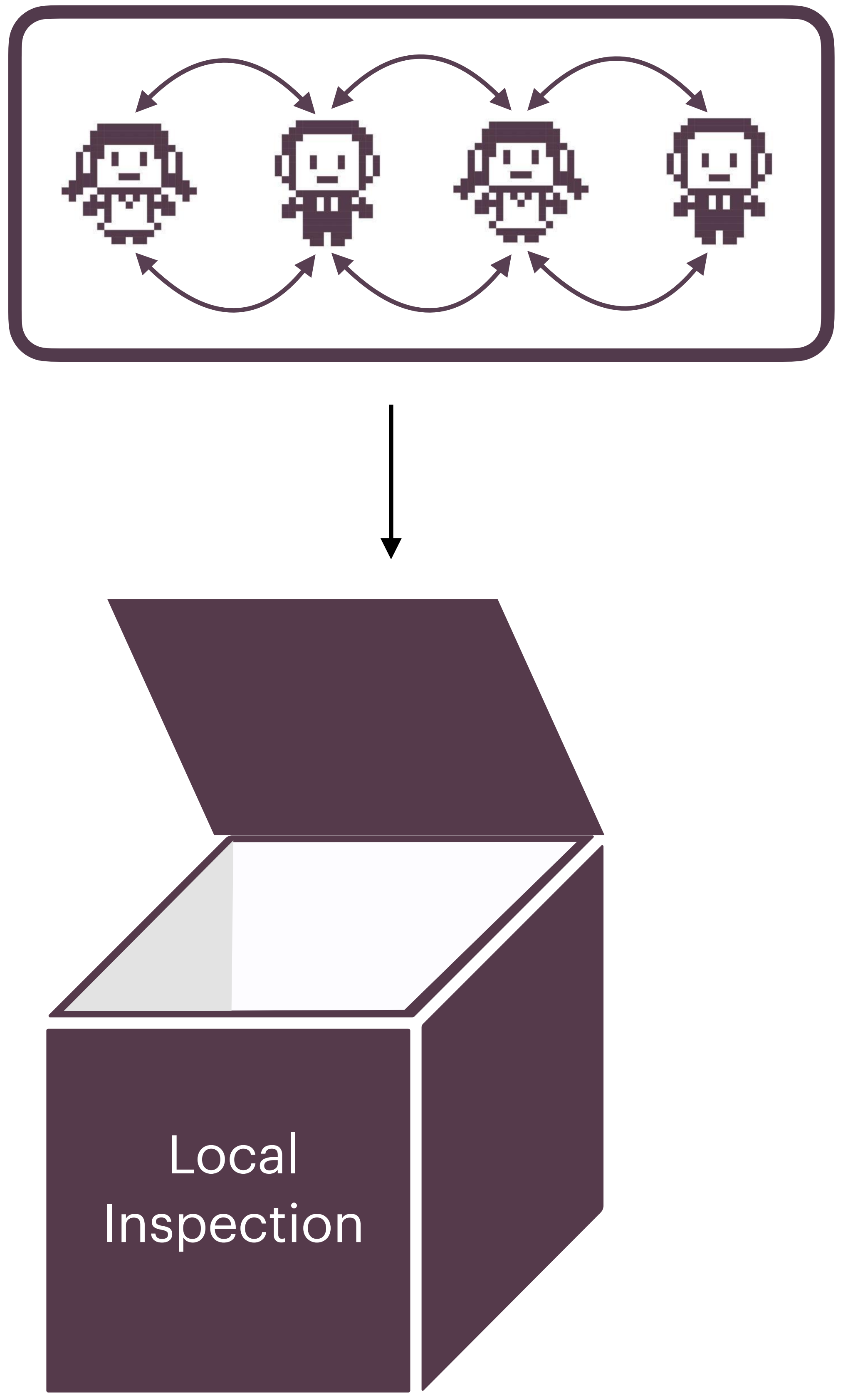
Vasconcelos and Zou et al 2024

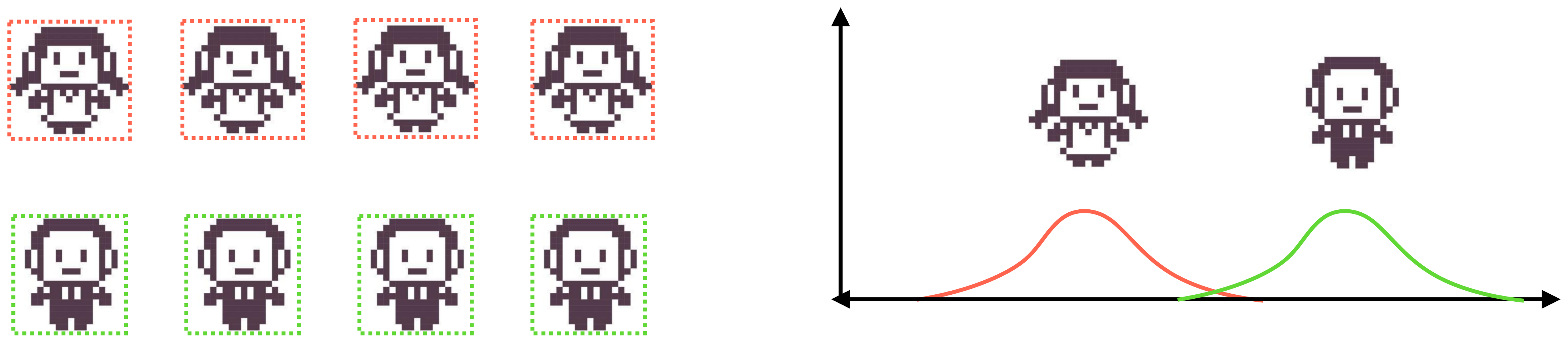
Local inspections take the form of verifying whether <u>relevant and known</u> patterns of human behavior appear in the simulation at the level of agents.

Vasconcelos and Zou et al 2024



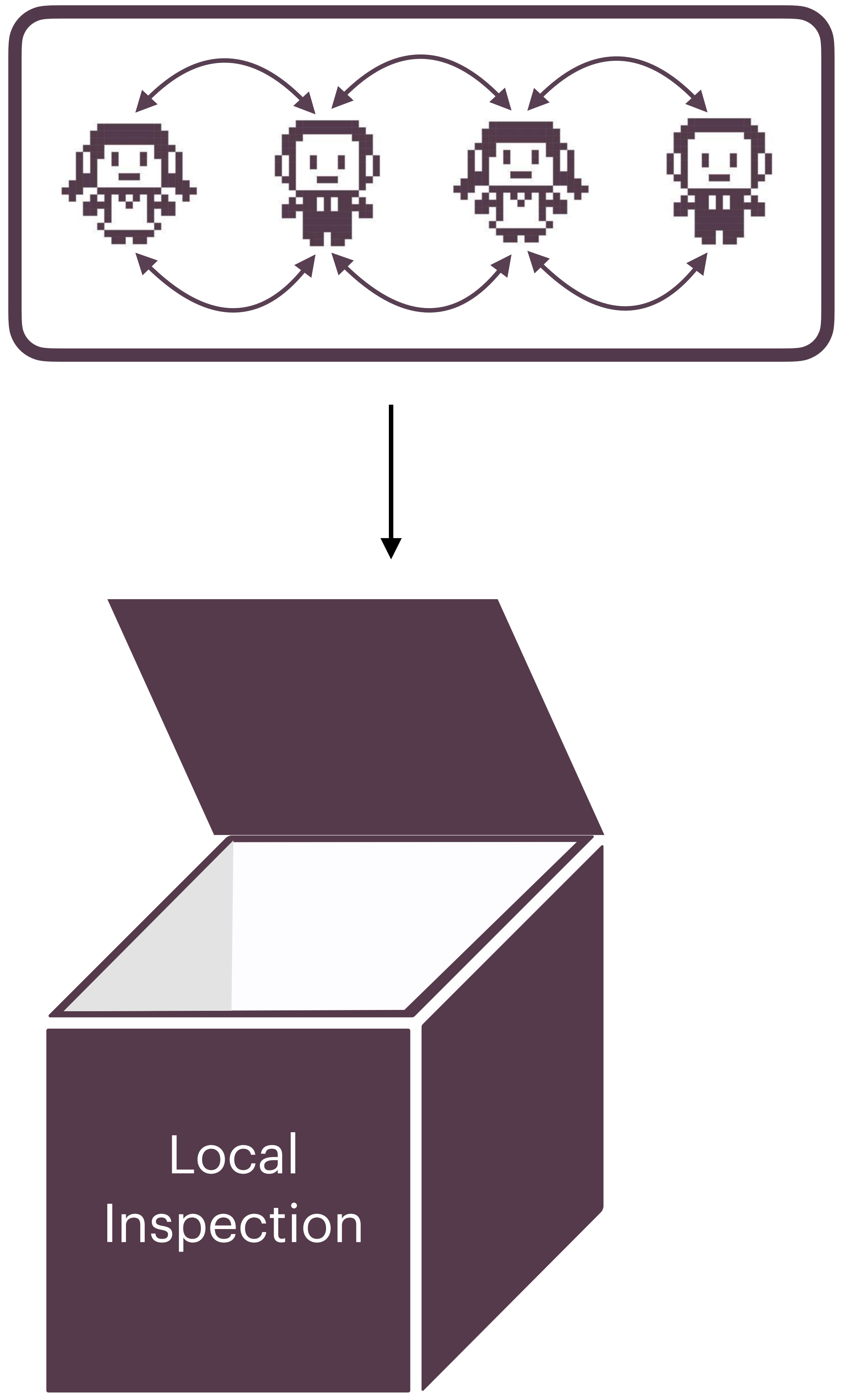
Vasconcelos and Zou et al 2024

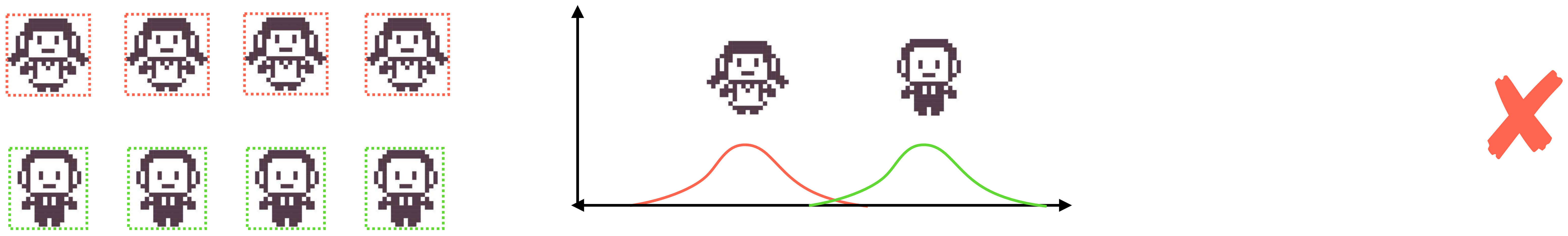




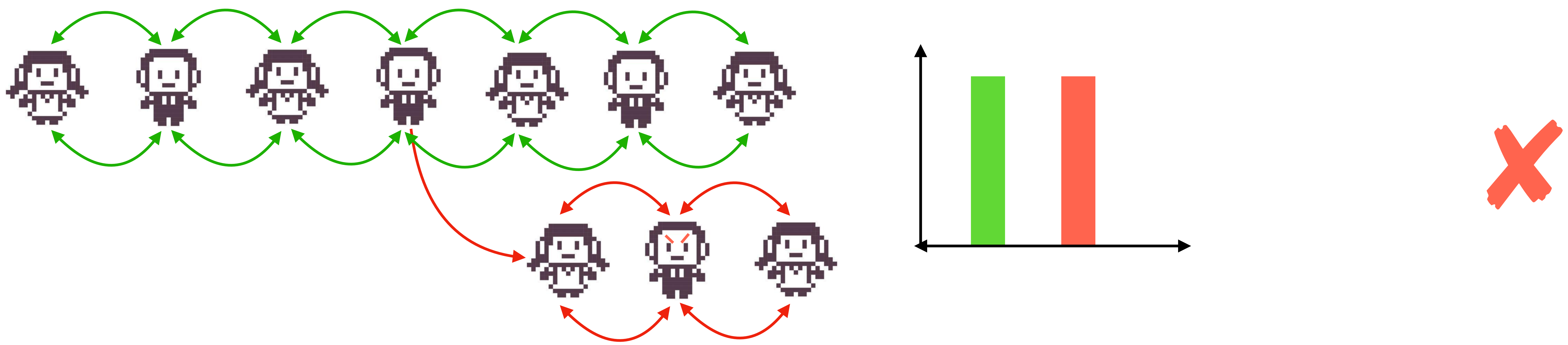
Reject if gender identity determines the main outcome with no strong explanatory theory

Vasconcelos and Zou et al 2024



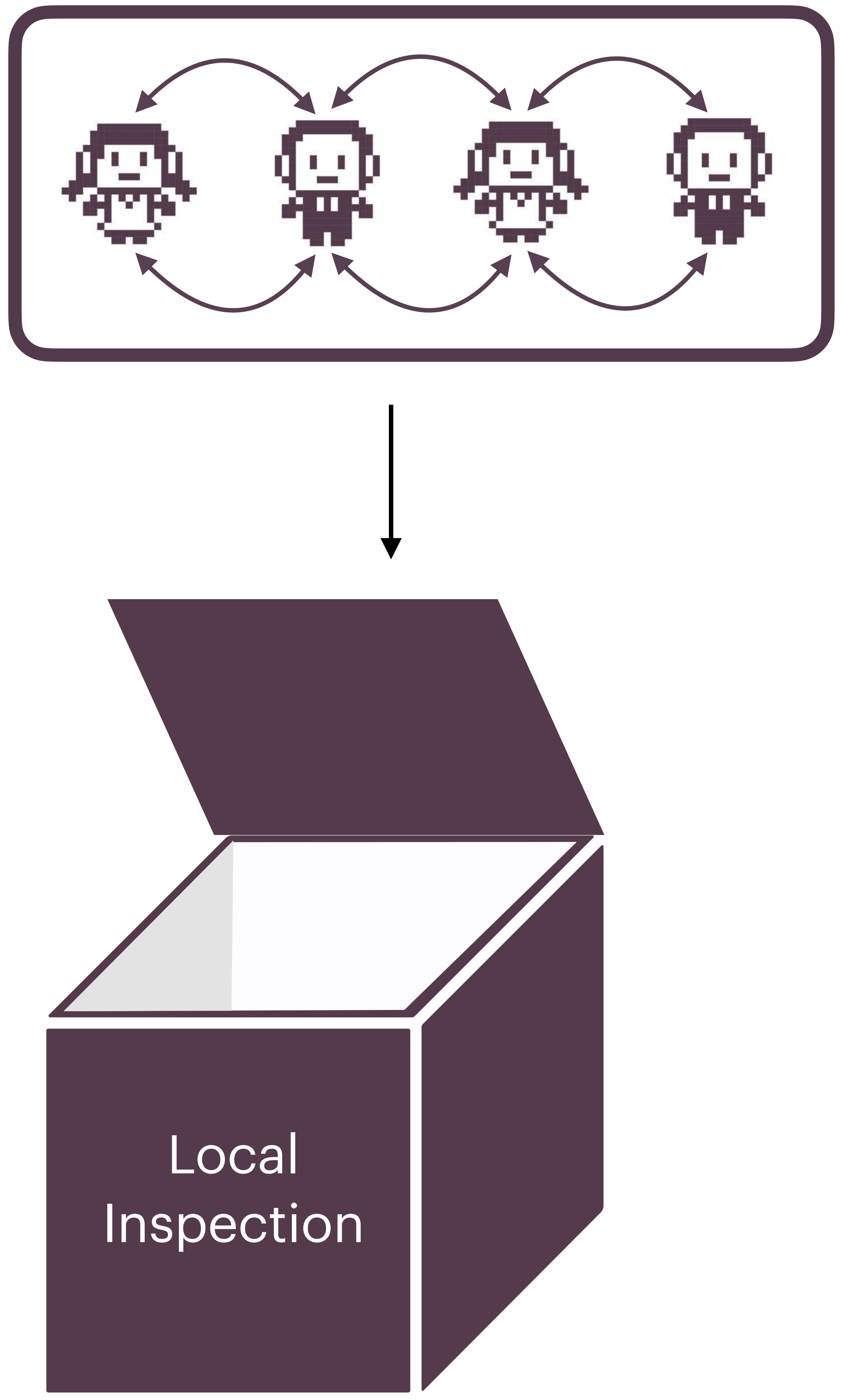


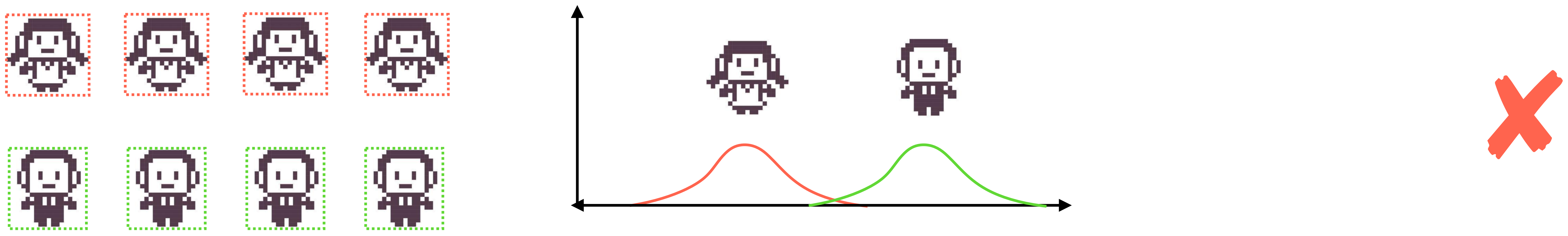
Reject if gender identity determines the main outcome with no strong explanatory theory



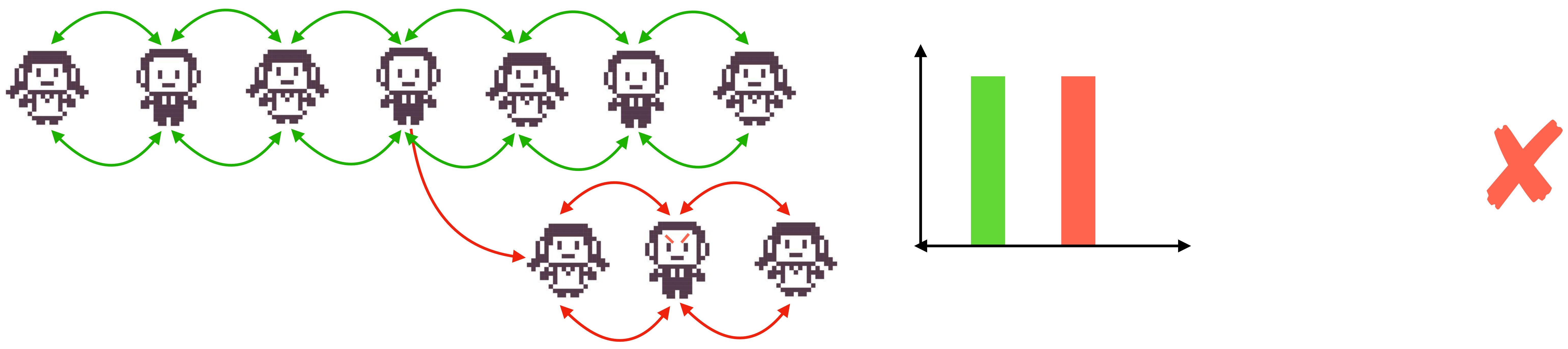
Reject if introducing highly disruptive agents does not cause changes in other agents’ behavior

Vasconcelos and Zou et al 2024





Reject if gender identity determines the main outcome with no strong explanatory theory



Reject if introducing highly disruptive agents does not cause changes in other agents' behavior

Reject if agents exhibit poor diversity-of-thought and exhibit unnaturally repetitive behavior

Vasconcelos and Zou et al 2024

Other examples of validation checks to include:

- Ensuring that certain cognitive biases are replicated
- Ensuring that increasing inflammatory rhetoric increases immediate polarization
- Ensuring that social contagion is found
- ...

Vasconcelos and Zou et al 2024

Local inspection allows practitioners to select relevant mechanisms, but applies them for <u>validation rather than direct incorporation.</u>

Ensuring the presence of these primitives can support trust while allowing for latent factors.

Vasconcelos and Zou et al 2024

But it's nearly impossible to check for all behaviors!

This is when researcher's should use their discretion. Just as when we run rigorous laboratory studies and check for as many confounding variables, we can do this with our simulations as well!

Vasconcelos and Zou et al 2024

Ultimately, the field is still figuring out how to do validation!

There are and will be many proposed methods for a “science" for LLM-based simulations.

Methods have already been proposed — either explicitly or implicitly — such as doing global audits or using the AgentBank.

Lecture Roadmap:

LLM Training 🤔	Training Data 📈	Running Inference 🏃‍♀️	Validation ✔	Reliance 👀
How do the ways (e.g., RLHF) in which models are trained affect the behaviors of agents?	What have these models learned? From where? How does this limit the accuracy of our agents?	How does the stochasticity and memorization of models affect accuracy? How do the architectures of the agents affect things?	Given the outputs of simulations, how might we validate them? How do we know what to trust?	How much trust should we put into the results of simulations? What happens if we put too much trust?

Reliance

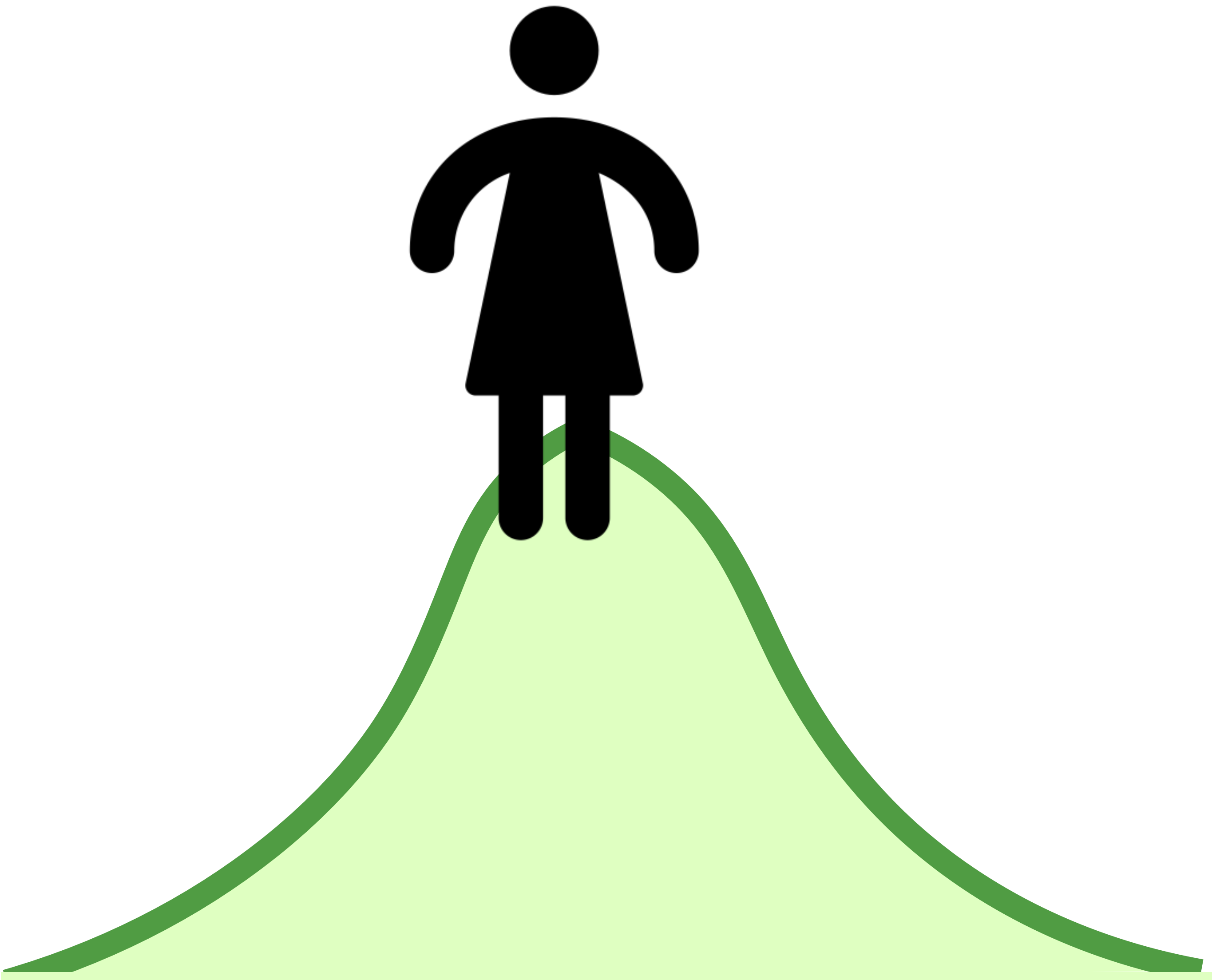
Reliance

After running our simulations and performing validation checks, how do we know when something is ready to be trusted? And what happens if we trust it too much?

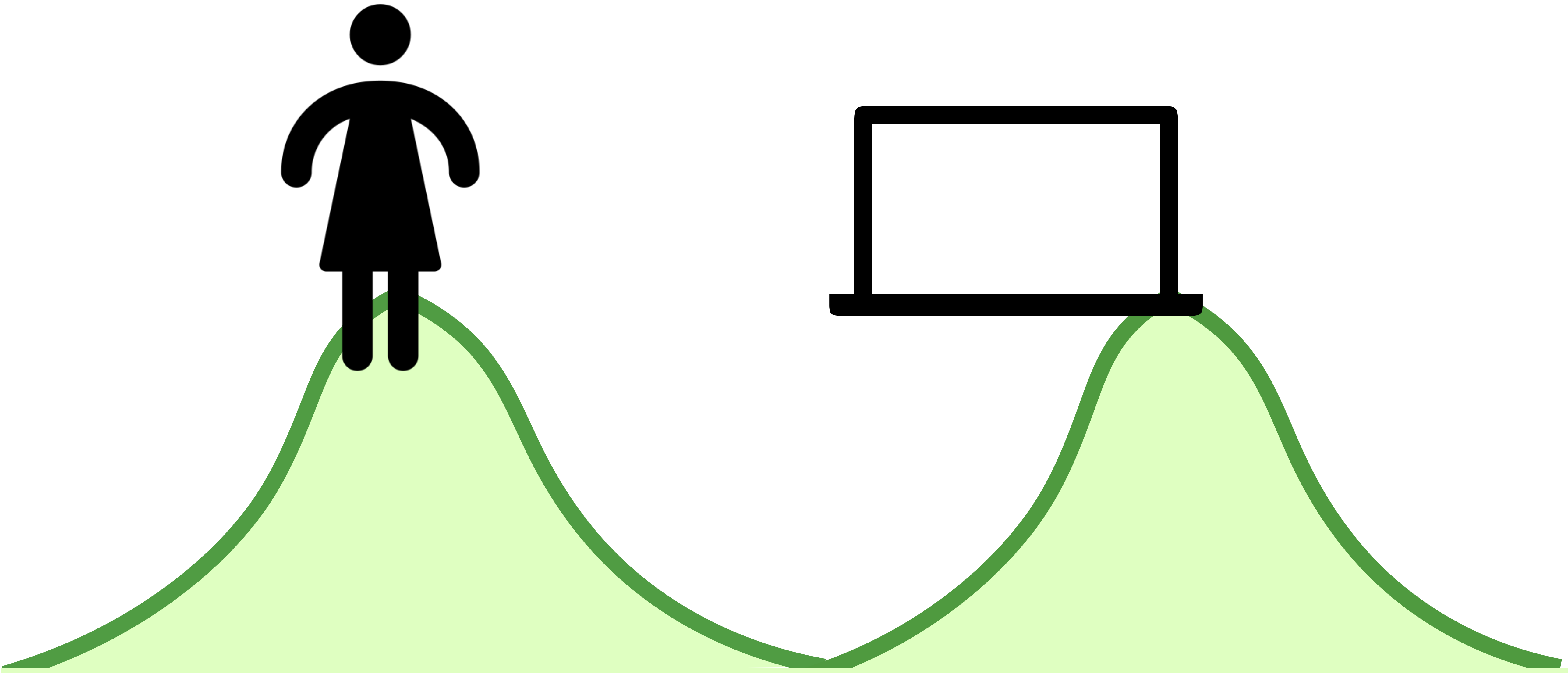
Reliance

After running our simulations and performing validation checks, how do we know when something is ready to be trusted? **And what happens if we trust it too much?**

This is already happening with other AI systems... a lot!



The goal: human-AI complementarity



The goal: human-AI complementarity

The goal: human-AI complementarity

But human-AI complementarity has not been realized

Well... why?

Overreliance:

Overreliance:

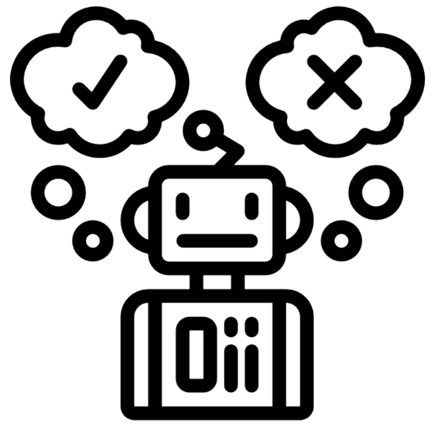
When people agree with an AI, even when the AI is wrong.

Overreliance:

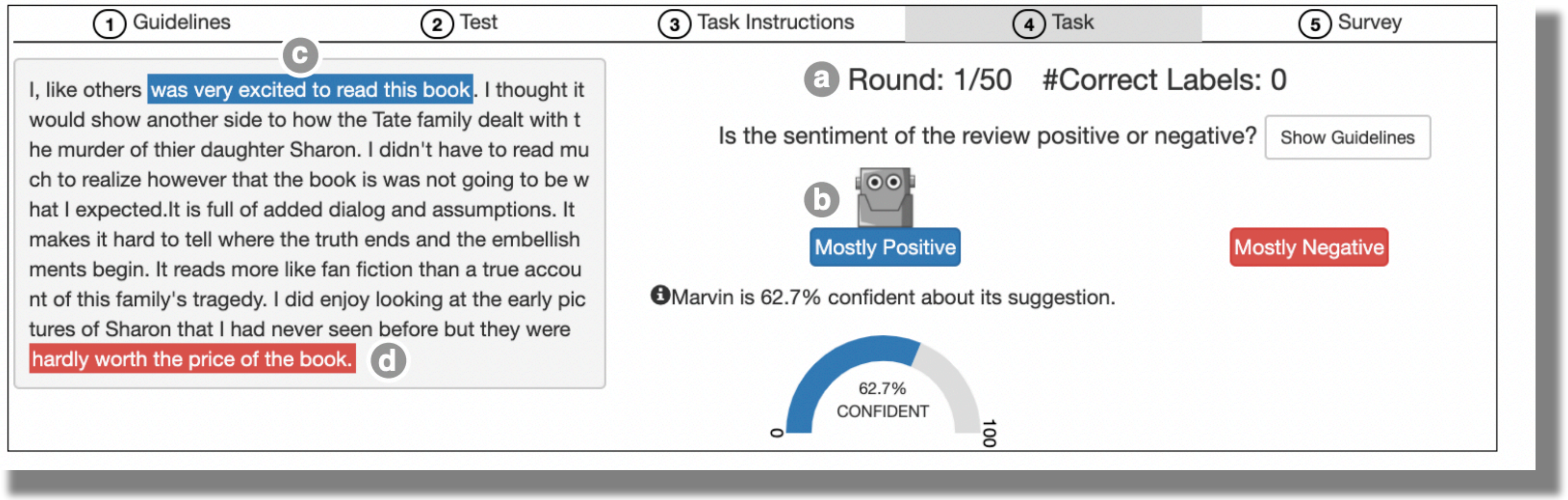
When people agree with an AI, even when the AI is wrong.

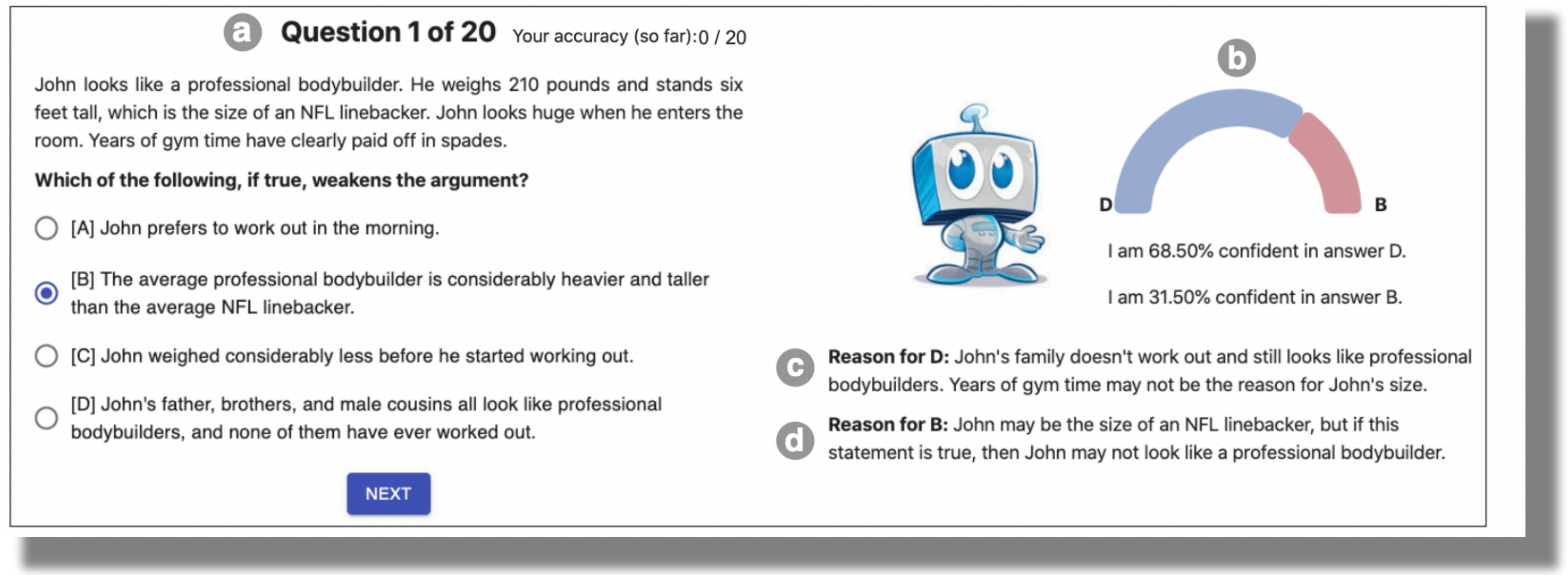
Human Decision-Maker

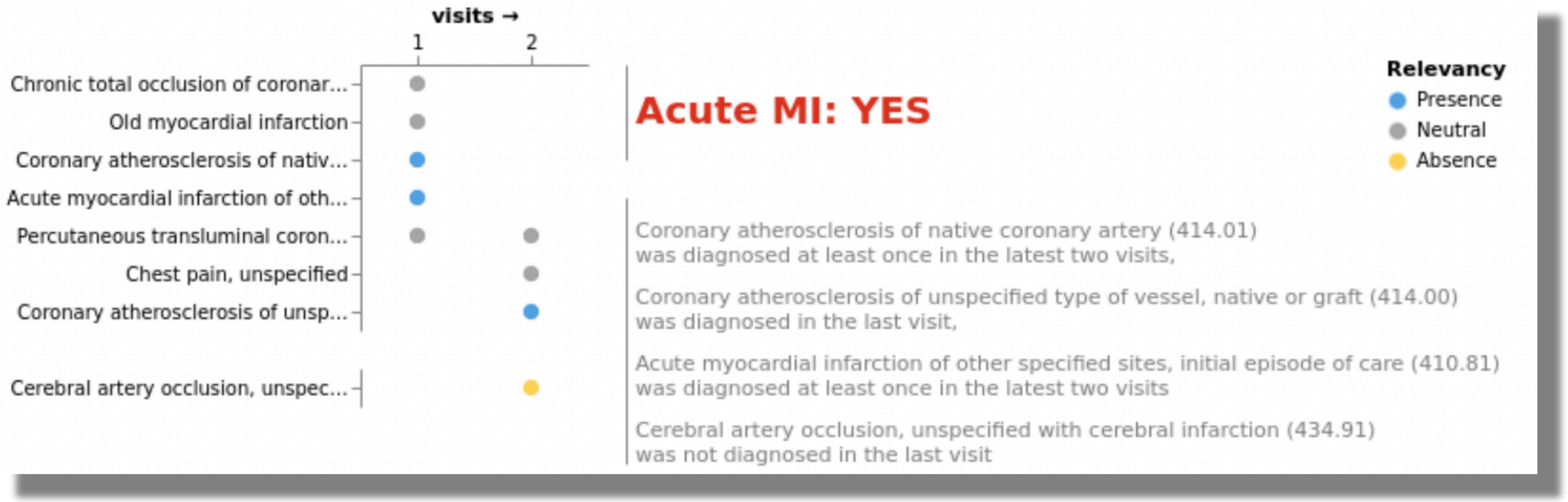
AI Agent	Reject AI’s Decision	Accept AI’s Decision
Correct	Underreliance	Appropriate Reliance
Incorrect	Appropriate Reliance	Overreliance

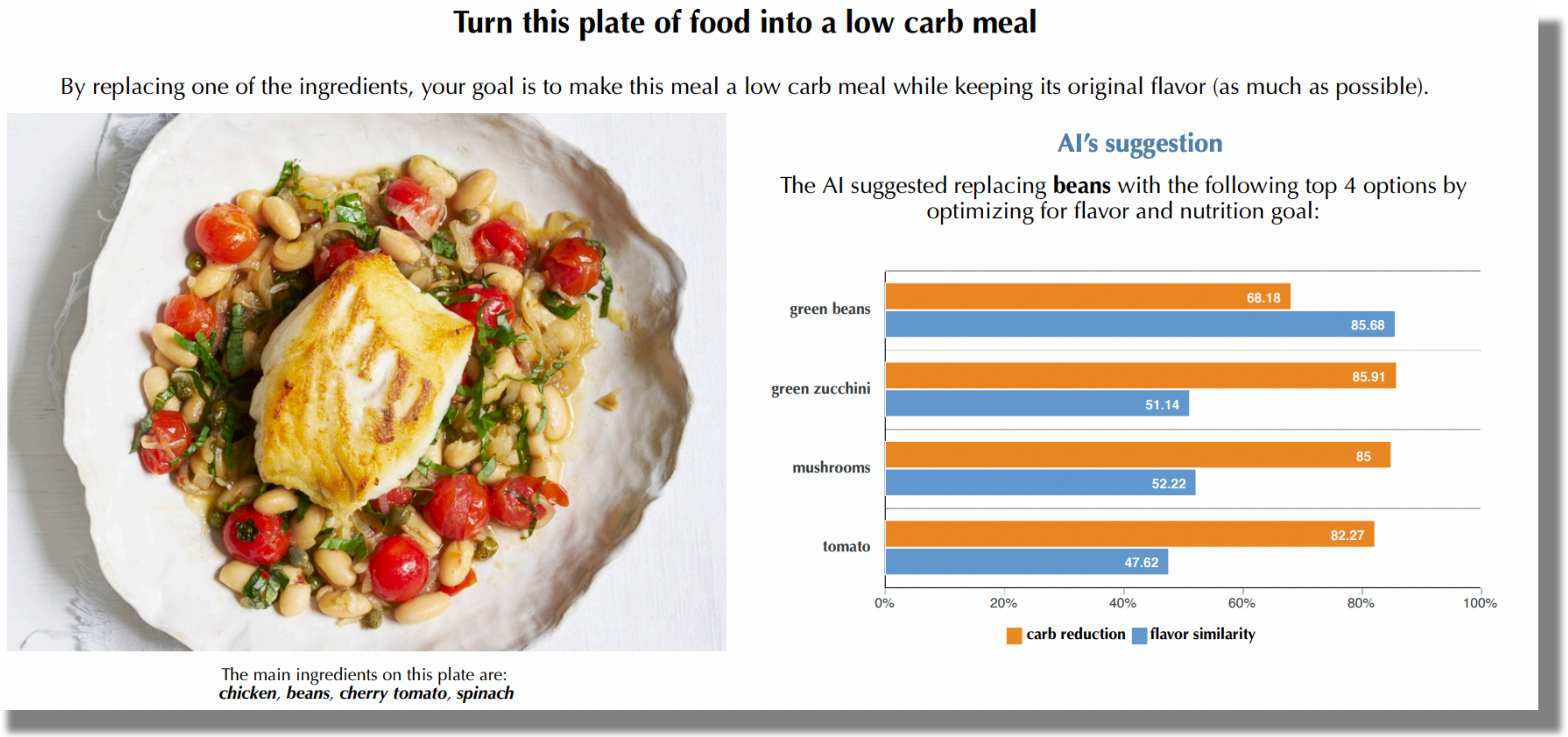


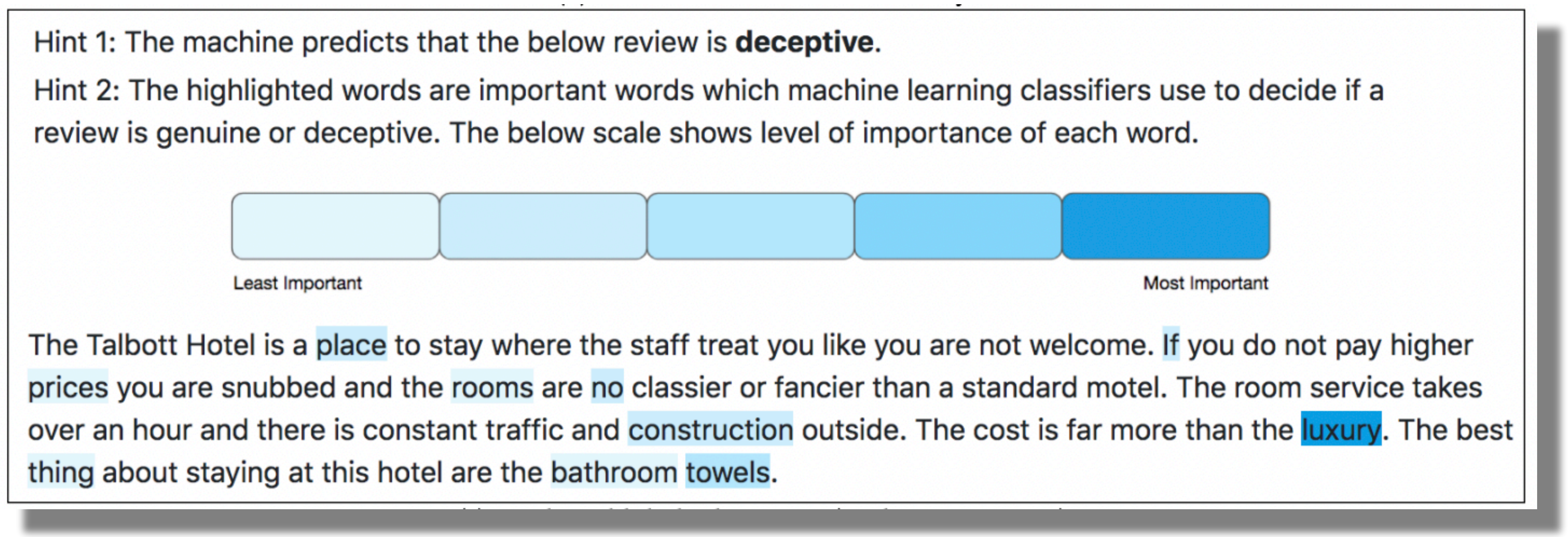
...this has been shown in a number of empirical studies!











Bansal, Wu, et al. 2021
Buçinca, et al. 2021
Lai, Tan 2019
Panigutti, Beretta, et al. 2022

…this has been shown in a number of empirical studies!

Making it easy to verify the AI (or alternatively, find errors in the model), through explanations or other means, will reduce overreliance.

* and this should be true of LLM-based simulations!

Vasconcelos et al., 2023

Reliance

Validation methods such as the one before help reduce the likelihood of overreliance, but prior work tells us that the errors need to be easy to verify! So, we still need HCI systems that allow us to perform whatever validation method, but to do so *easily!*

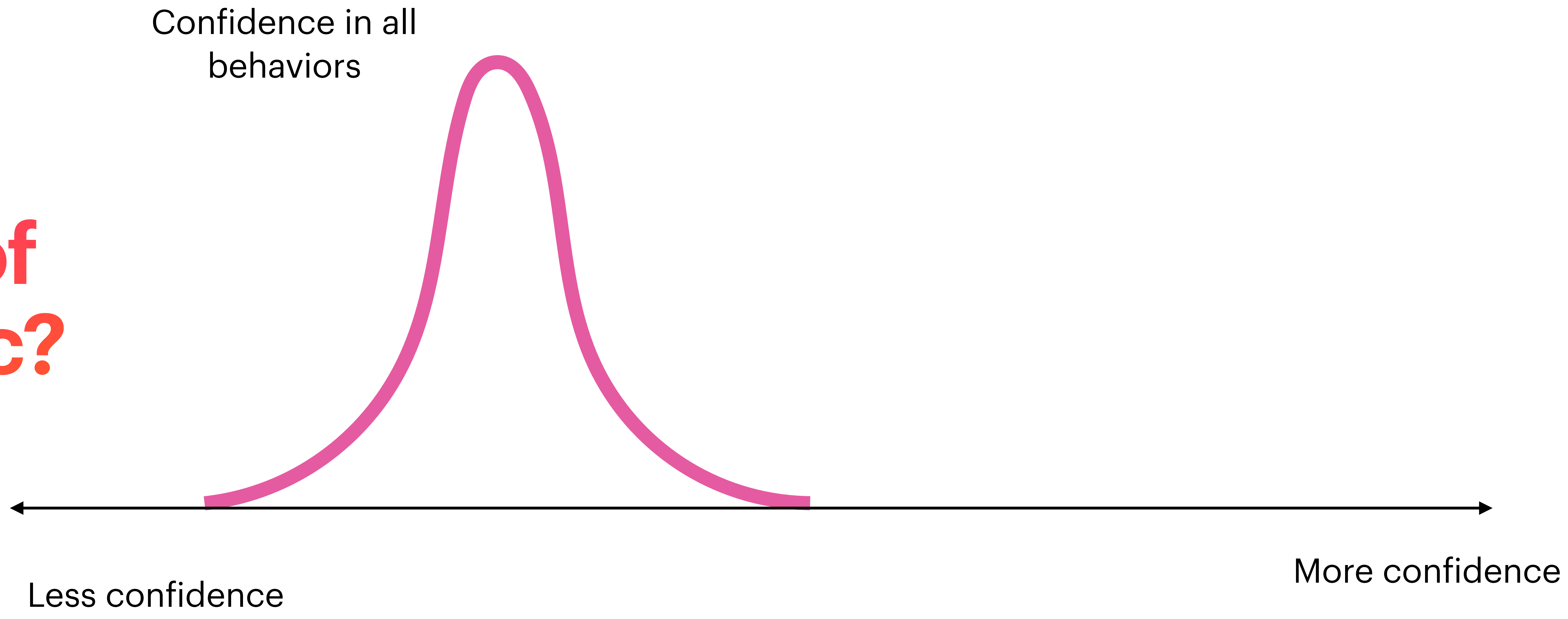
* open area!

Reliance

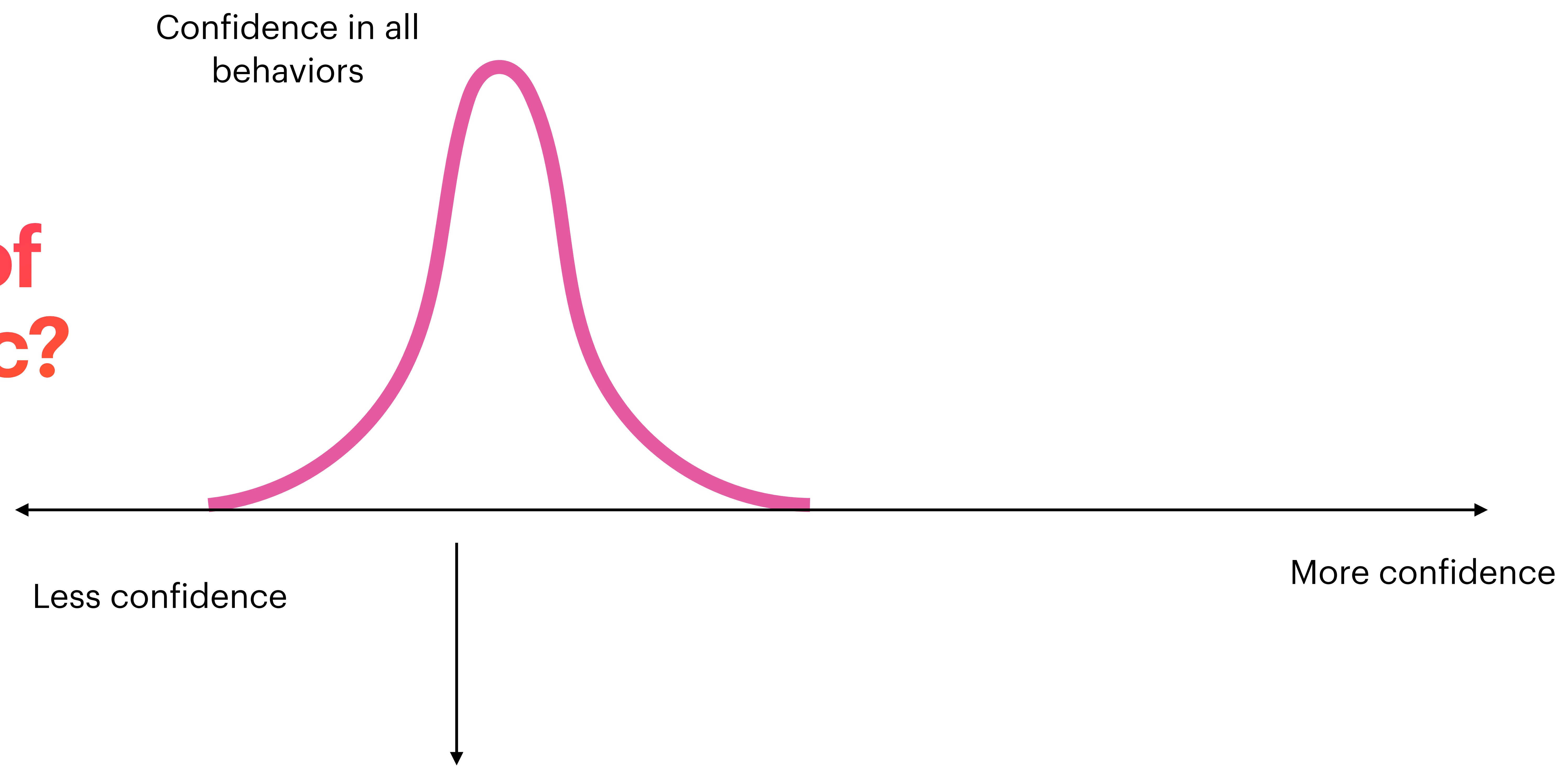
Validation methods such as the one before help reduce the likelihood of overreliance, but prior work tells us that the errors need to be easy to verify! So, we still need HCI systems that allow us to perform whatever validation method, but to do so *easily!*

But even with the validation methods, it's still unclear how much **epistemic confidence** you should be putting into the results of simulations...

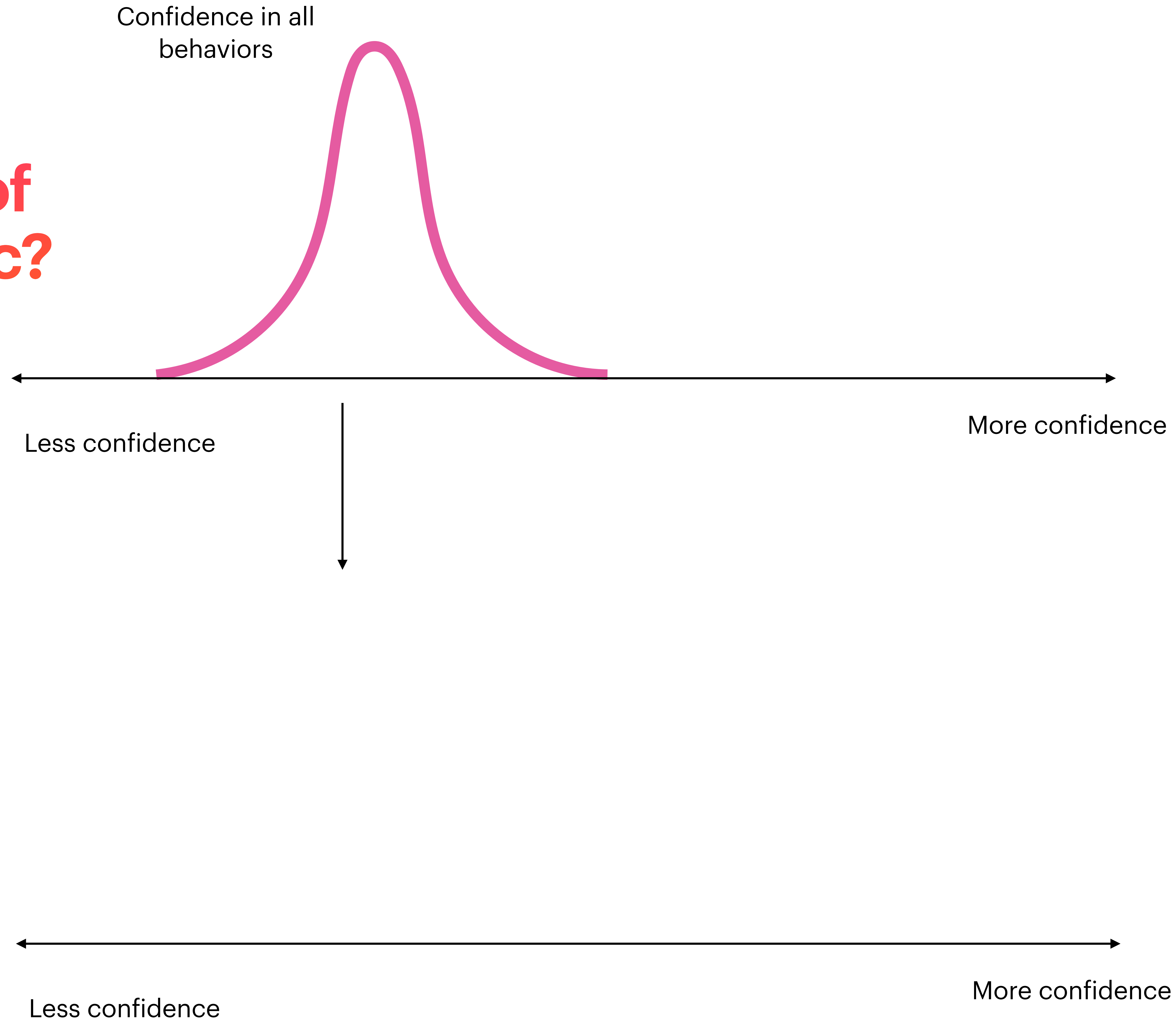
How much epistemic confidence should we have that the outcome of the simulation is realistic?



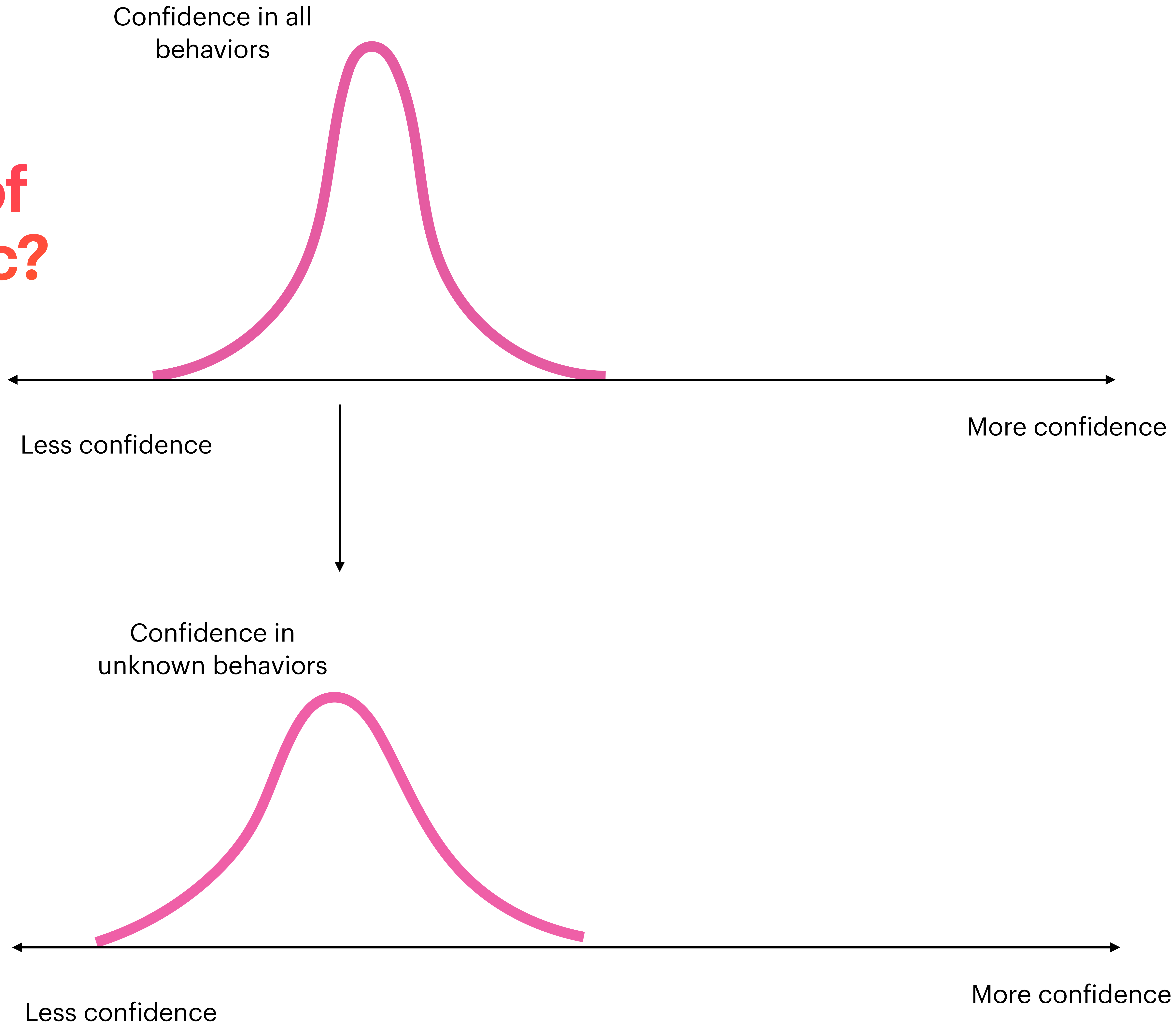
How much epistemic confidence should we have that the outcome of the simulation is realistic?



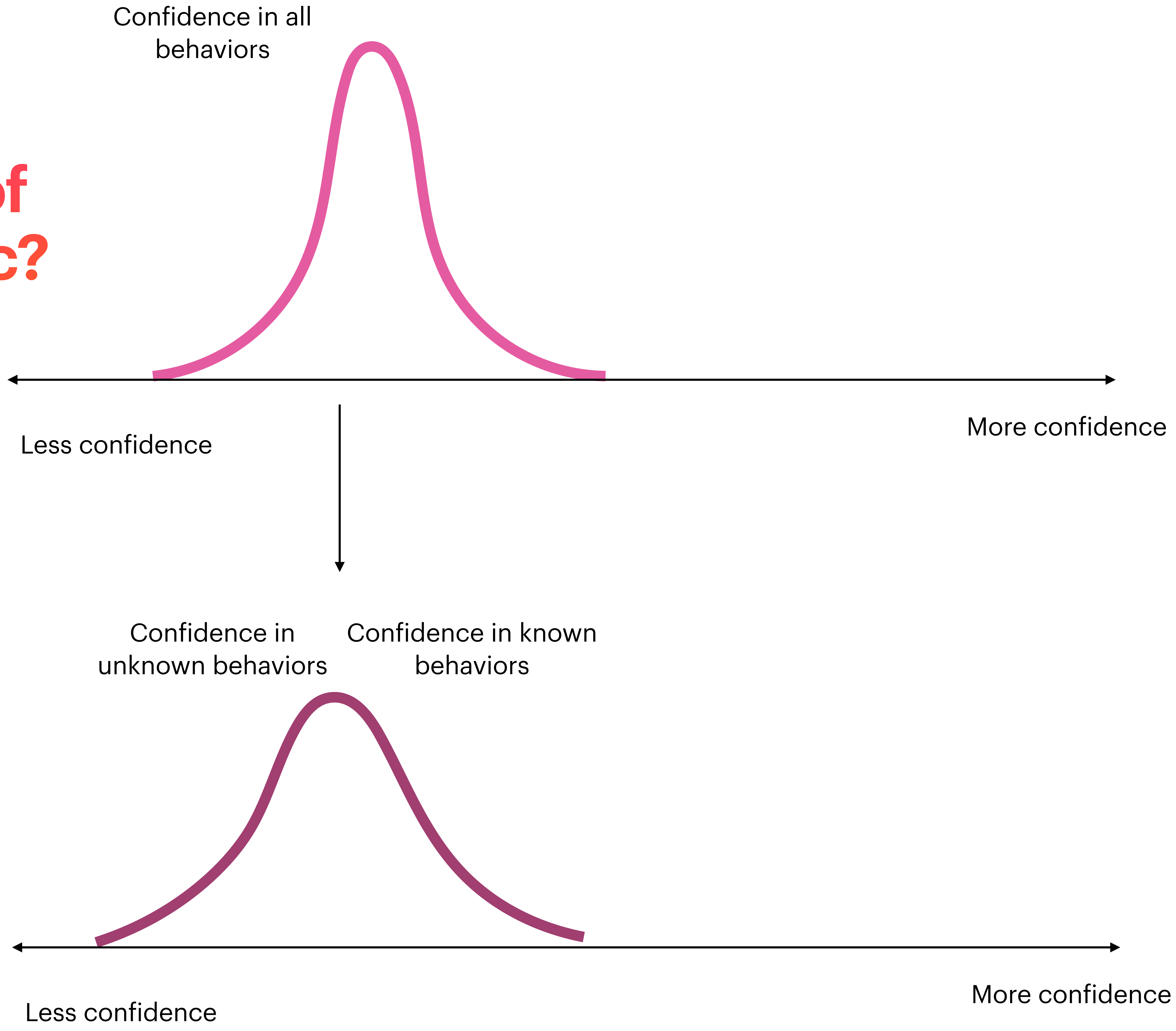
How much epistemic confidence should we have that the outcome of the simulation is realistic?



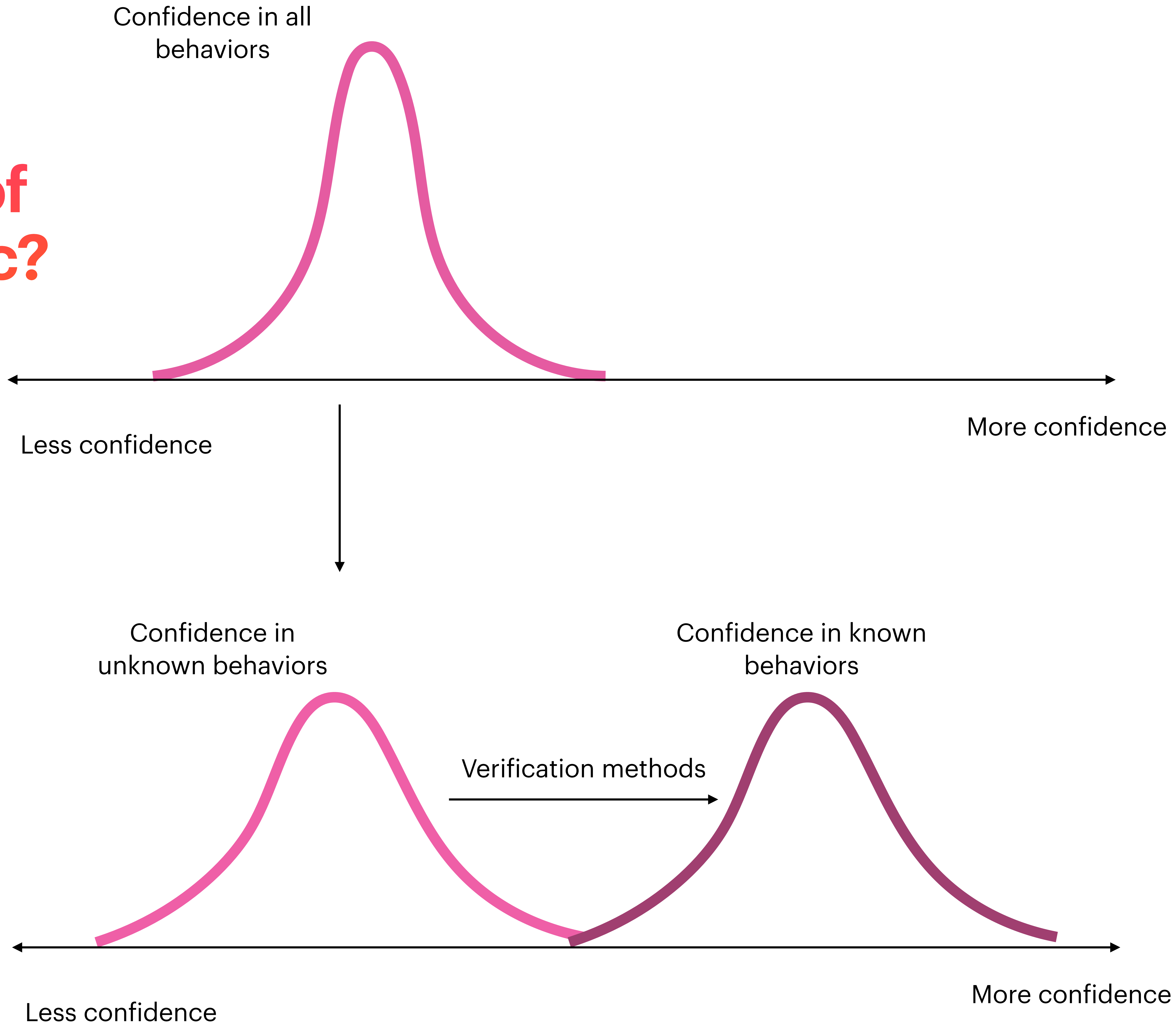
How much epistemic confidence should we have that the outcome of the simulation is realistic?



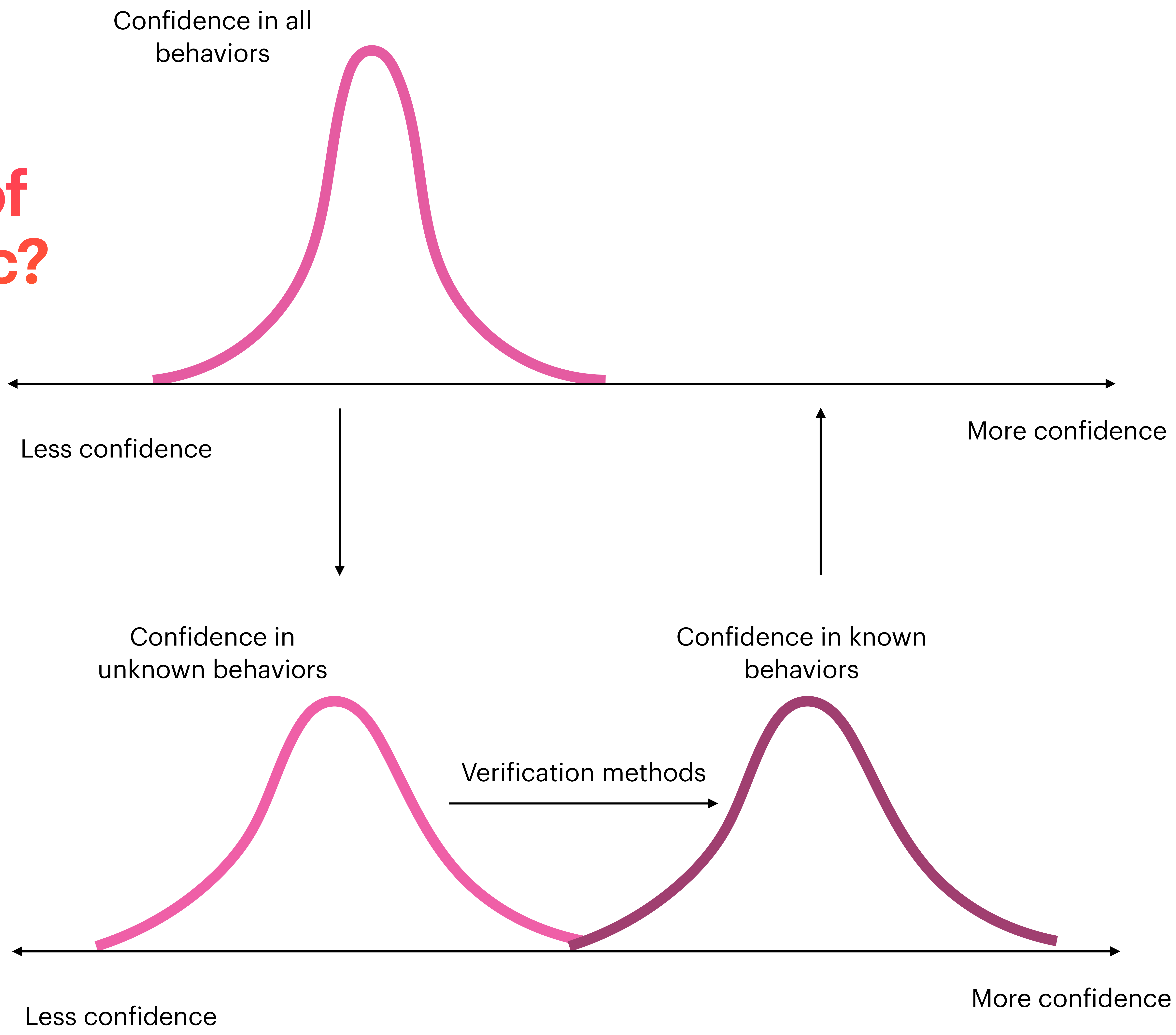
How much epistemic confidence should we have that the outcome of the simulation is realistic?



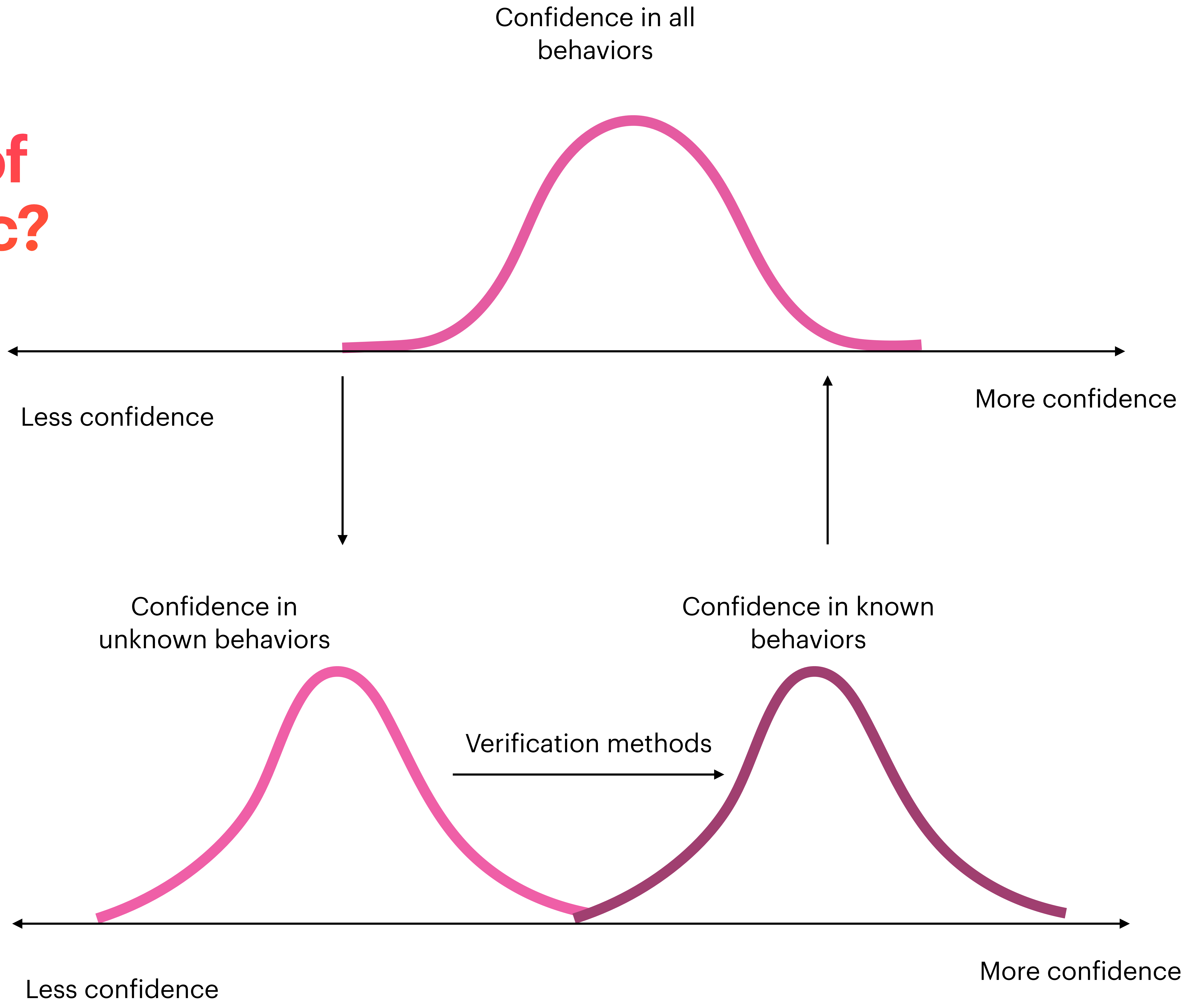
How much epistemic confidence should we have that the outcome of the simulation is realistic?



How much epistemic confidence should we have that the outcome of the simulation is realistic?



How much epistemic confidence should we have that the outcome of the simulation is realistic?



What kinds of applications are appropriate for using simulated outcomes?

Ideally, applications should be ones where a low level of confidence is sufficient and no alternative methods exist.

Requires low epistemic confidence	Requires high epistemic confidence
Hypothesis generation for feed algorithm changes	Measuring community resilience to toxicity

Reliance

But even with these guidelines and methods, simulations can still be **misused** and misinterpreted!

What kinds of applications are appropriate for using simulated outcomes?

Ideally, applications should be ones where a low level of confidence is sufficient and no alternative methods exist.

	Requires low epistemic confidence	**Requires high epistemic confidence**
Alternatives don't exist	Hypothesis generation for feed algorithm changes	Measuring community resilience to toxicity
Alternatives are prohibitive	Exploration prior to user interviews	Election forecasting
Alternatives exist	Testing content moderation changes	Participatory design methods

Discourse about this very question in the community...

Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study

Perttu Hämäläinen*
perttu.hamalainen@aalto.fi
Aalto University
Espoo, Finland

Mikke Tavast*
mikke.tavast@aalto.fi
Aalto University
Espoo, Finland

Anton Kunnari
anton.kunnari@helsinki.fi
University of Helsinki
Helsinki, Finland

ABSTRACT

Collecting data is one of the bottlenecks of Human-Computer Interaction (HCI) research. Motivated by this, we explore the potential of large language models (LLMs) in generating synthetic user research data. We use OpenAI's GPT-3 model to generate open-ended questionnaire responses about experiencing video games as art, a topic not tractable with traditional computational user models. We test whether synthetic responses can be distinguished from real responses, analyze errors of synthetic data, and investigate content similarities between synthetic and real data. We conclude that GPT-3 can, in this context, yield believable accounts of HCI experiences. Given the low cost and high speed of LLM data generation, synthetic data should be useful in ideating and piloting

AI: *My favorite writing tool would have to be Microsoft Word. I find it very user-friendly and versatile, and it has everything I need to get the job done.*

Researcher: Why do you prefer Word over LaTeX?

AI: *LaTeX is a great tool for typesetting, but I find it to be much more difficult to use than Word. With Word, I can easily format my document the way I want it to look, and I don't have to worry about code.*

The dialogue above was generated using OpenAI Playground[1], a tool that allows one to input a piece of text—*a prompt*—and ask the GPT-3 large language model (LLM) [10] to generate a plausible continuation. We wrote the boldface parts and let GPT-3 generate the italicized continuations. The result is characteristic of the

The Illusion of Artificial Inclusion

William Agnew
Carnegie Mellon University
Pittsburgh, PA, USA

A. Stevie Bergman
Google DeepMind
New York, NY, USA

Jennifer Chien
University of California San Diego
San Diego, CA, USA

Mark Díaz
Google Research
New York, NY, USA

Seliem El-Sayed
Google DeepMind
London, UK

Jaylen Pittman
Stanford University
Stanford, CA, USA

Shakir Mohamed
Google DeepMind
London, UK

Kevin R. McKee
Google DeepMind
London, UK

ABSTRACT

Human participants play a central role in the development of modern artificial intelligence (AI) technology, in psychological science, and in user research. Recent advances in generative AI have attracted growing interest to the possibility of replacing human participants in these domains with AI surrogates. We survey several such "substitution proposals" to better understand the arguments for and against substituting human participants with modern gen-

1 INTRODUCTION

Participation is a foundational element of the social-behavioral sciences and in the design of new technology. In psychology, user research, human-computer interaction (HCI), and other related fields, research participants offer a window into human cognition and decision making. In the development of new technologies, human participants ground the design process in real-life needs, perspectives, and experiences.

Reliance

But even with these guidelines and methods, simulations can still be misused and **misinterpreted**!

Reliance

After validation methods have been made and guidelines on epistemic confidence set, there is still a big risk that (1) these simulations are purposefully used by people in ways that justify unethical ends, (2) but even when trying to use simulations in good faith, researchers, policy makers, industry professionals get the wrong insights from simulations.

Reliance

After validation methods have been made and guidelines on epistemic confidence set, there is still a big risk that **(1) these simulations are purposefully used by people in ways that justify unethical ends,** (2) but even when trying to use simulations in good faith, researchers, policy makers, industry professionals get the wrong insights from simulations.

Reliance

After validation methods have been made and guidelines on epistemic confidence set, there is still a big risk that (1) these simulations are purposefully used by people in ways that justify unethical ends, **(2) but even when trying to use simulations in good faith, researchers, policy makers, industry professionals get the wrong insights from simulations.**

The believability of agents poses new sociotechnical risks, via interpretive errors

Here are some things you can do:

Pick the right level of abstraction

Perturb design decisions to understand causality

Use human cognition metaphors with purpose

Track data provenance

Lecture Roadmap:

LLM Training 🤔	Training Data 📈	Running Inference 🏃‍♀️	Validation ✔	Reliance 👀
How do the ways (e.g., RLHF) in which models are trained affect the behaviors of agents?	What have these models learned? From where? How does this limit the accuracy of our agents?	How does the stochasticity and memorization of models affect accuracy? How do the architectures of the agents affect things?	Given the outputs of simulations, how might we validate them? How do we know what to trust?	How much trust should we put into the results of simulations? What happens if we put too much trust?

In summary…

Lots of limitations but also lots of opportunities in…

- Modeling
- Data
- Inference
- Architecture
- Validation
- Tools for reliance
- Use