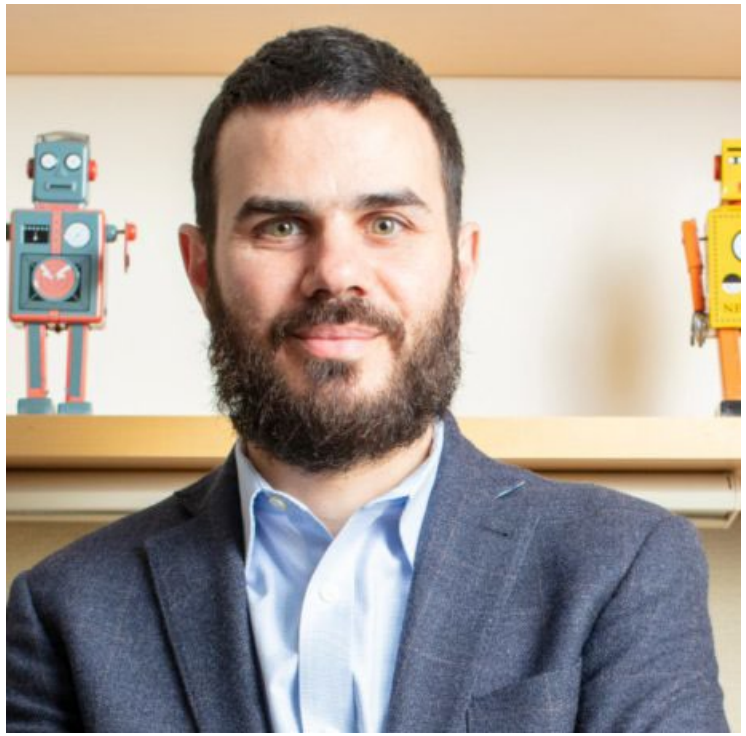
Integrating Generative AI into Experiments

Thomas H. Costello

Assistant Professor of Psychology, American University



Dave Rand
MIT

Gordon Pennycook
Cornell University

Hause Lin
MIT

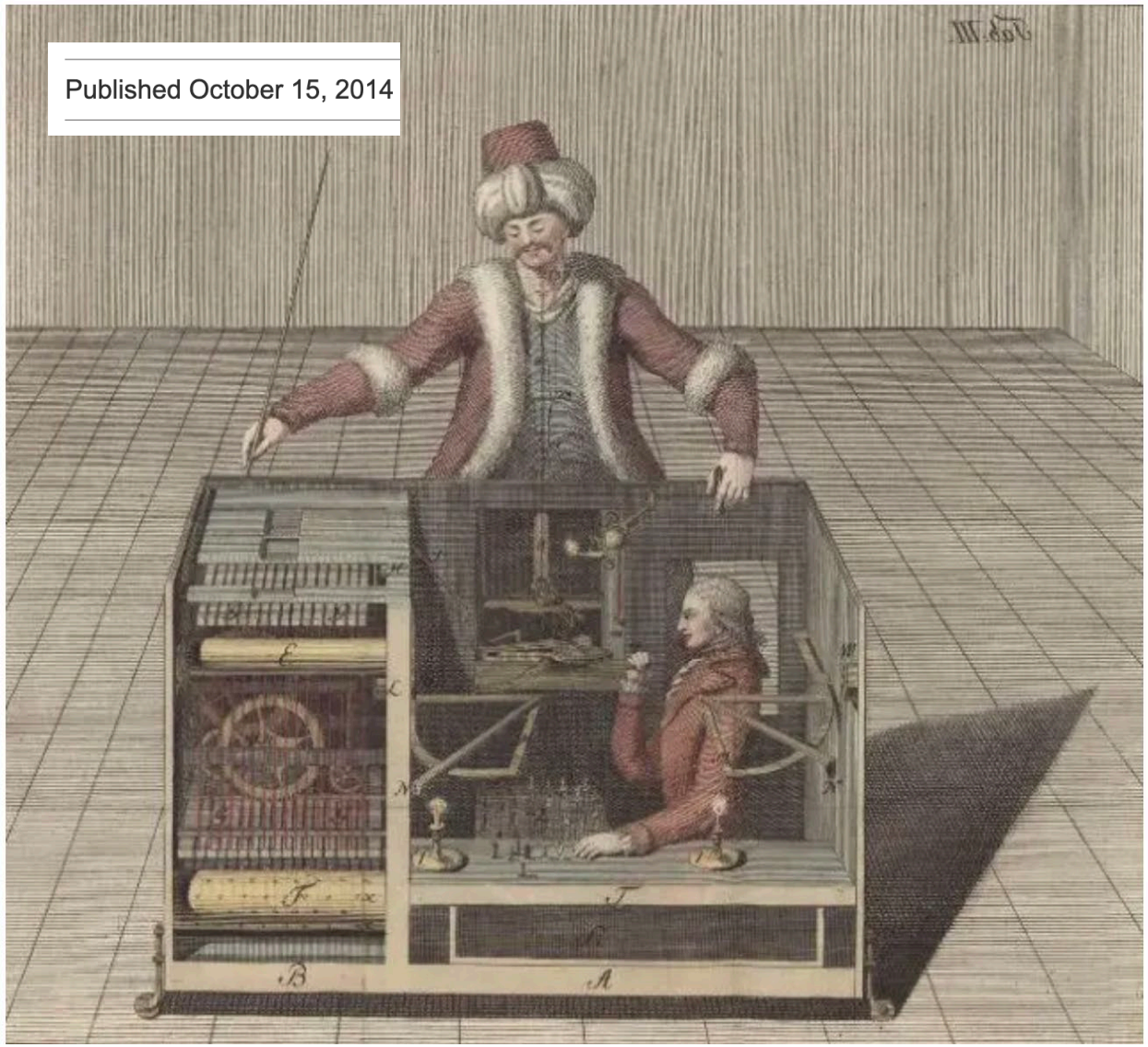
ADVANTAGES: Flexible, adaptive, conducive to creativity, ecologically valid, minimal fraud, hypothesis-generative

PROBLEMS: Not standardized, EXPENSIVE, difficult to replicate, most information not recorded, limited pool of subjects

Mechanical Turk: The New Face of Behavioral Science?

Priceonomics

Published October 15, 2014



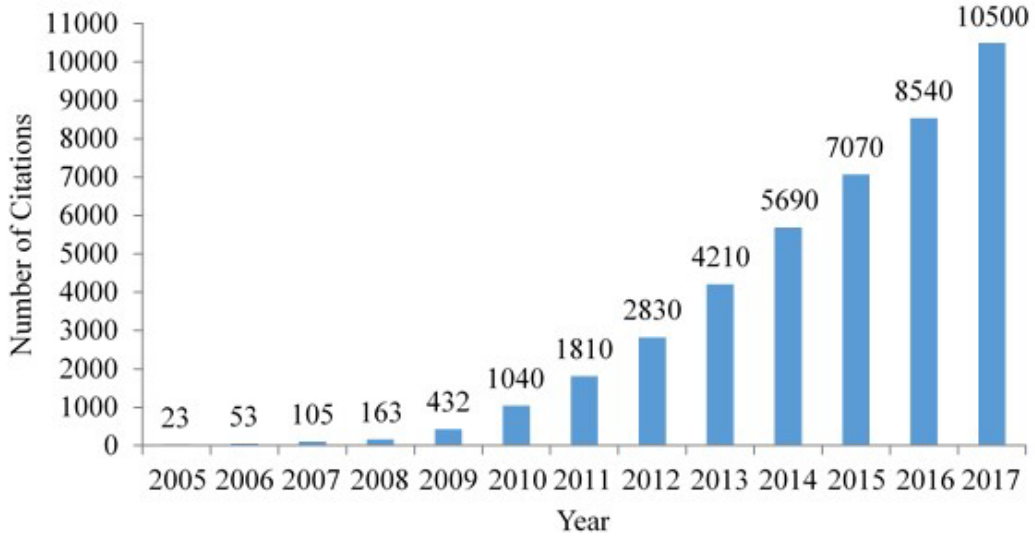


Figure 1. Number of Citations for "Mechanical Turk" in Google Scholar by Year

Source: https://societyforpsychotherapy.org/an-mturk-primer-for-psychotherapy-researchers/

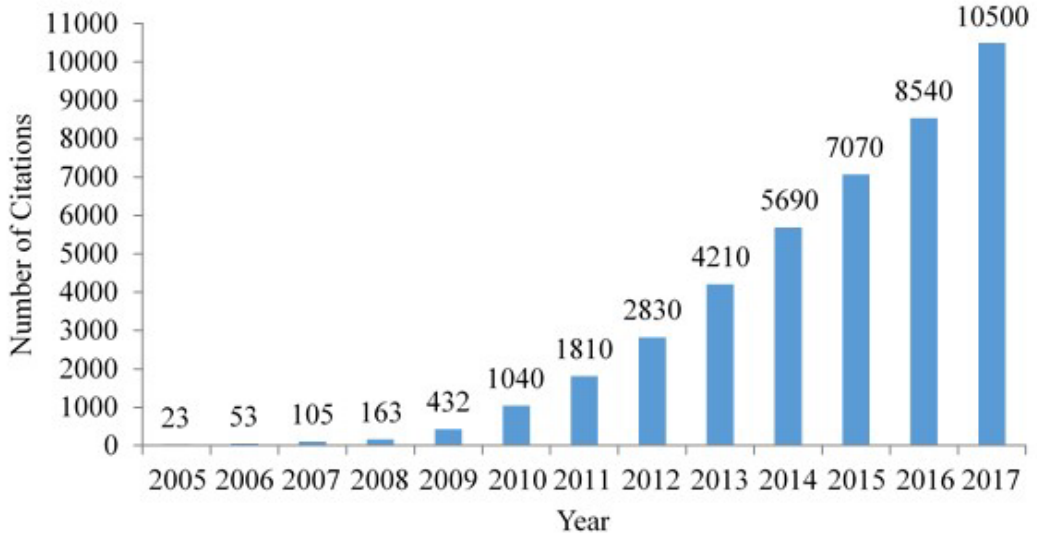
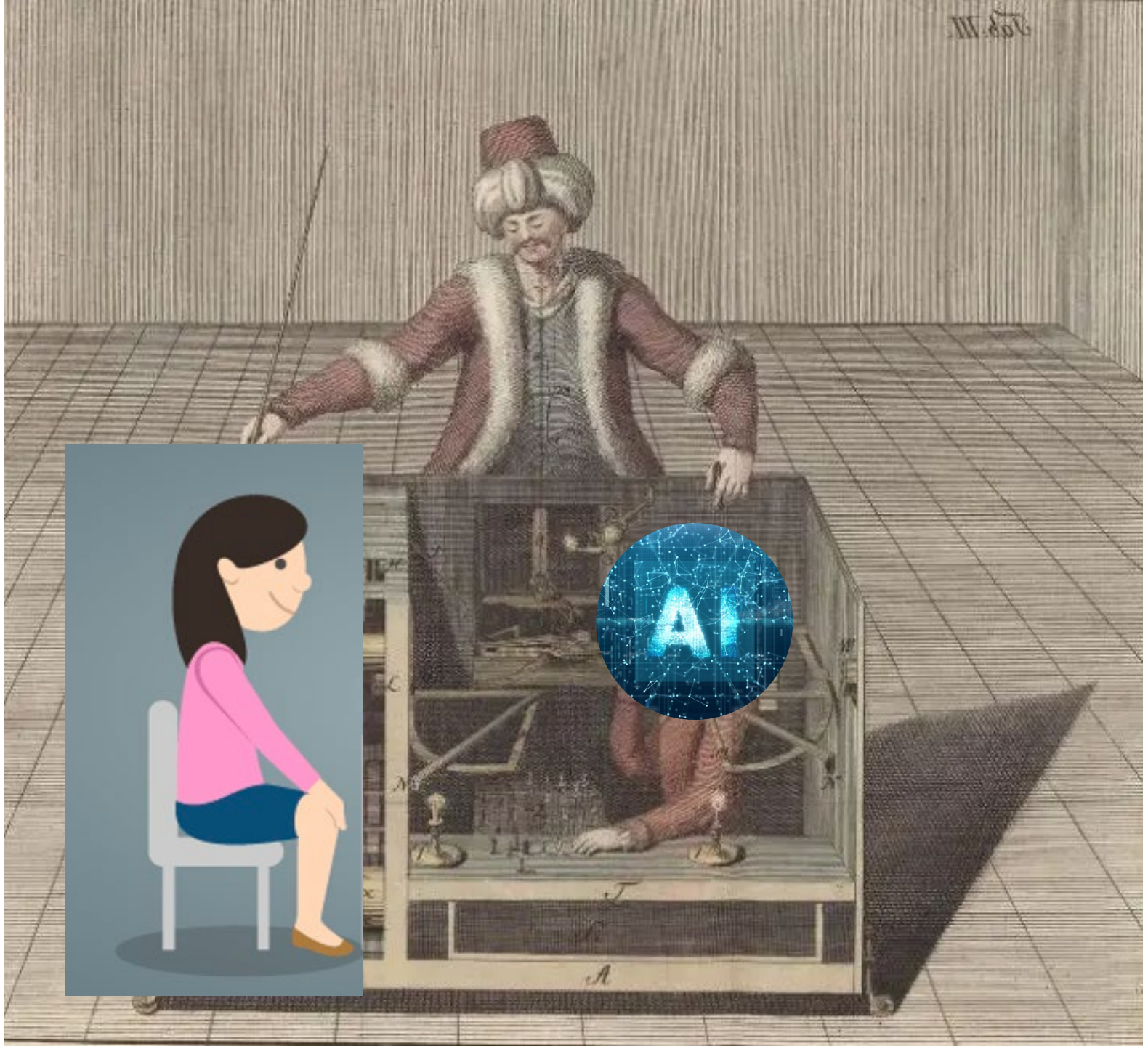


Figure 1. Number of Citations for "Mechanical Turk" in Google Scholar by Year

ADVANTAGES: Scalable, inexpensive, reach many populations, easy to replicate, easy to iterate

PROBLEMS: impersonal, inflexible, blunt; prone to low quality participants (bot, inattention), boring (low engagement/focus)

Tab. III.
E
C
F
N
T
B
A


AI

The third wave of behavioral science?

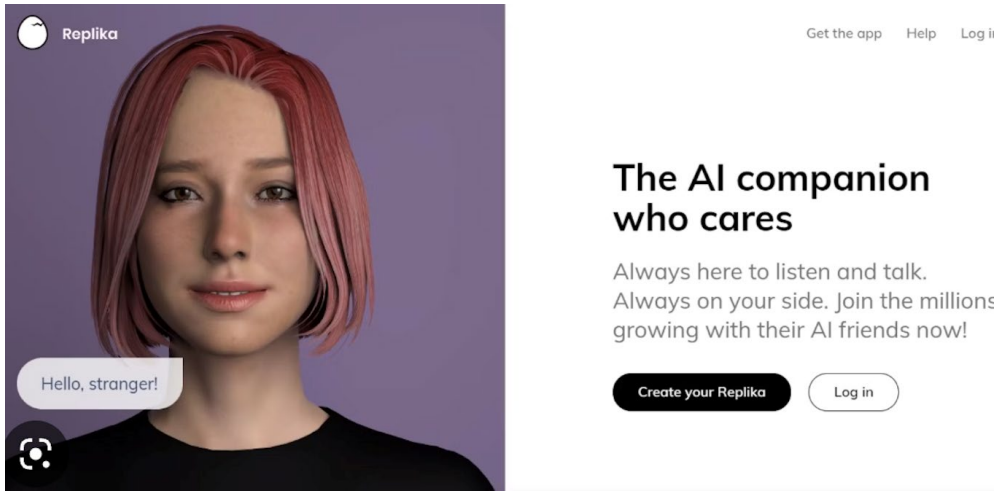
The third wave of behavioral science?

- **Bringing AI (LLMs, etc) into the lab** – to act as a “confederate+”
- AI models can :
  - Deliver treatments
    - Responsive to participants’ state
    - Variable along pre-specified axes, standardized on others
  - Monitor for bots
  - Determine appropriate group assignments
  - Generate and collect feedback
  - Much more

- A new kind of ecological validity

The third wave of behavioral science?

- **Bringing AI (LLMs, etc) into the lab** – to act as a “confederate+”
- AI models can :
  - **Deliver treatments**
    - Responsive to participants’ state
    - Variable along pre-specified axes, standardized on others
  - **Monitor for bots**
  - Determine appropriate group assignments
  - Generate and collect feedback
  - Much more
- A new kind of ecological validity



Advantages	AI-to-person
Flexible/adaptive	✓
Conducive to creativity	✓
(More) Ecologically valid	✓
Minimal participant fraud	✓
Hypothesis-generative	✓
Relatively engaging	✓
Scalable	✓
Inexpensive	✓
Reach many populations	✓
Easy to replicate	✓
Easy to iterate	✓
Standardized (at level of "prompt”	✓
Most information recorded	✓

What this looks like, roughly, in emerging work

- Subject enters study
- LLM monitors some/all of their answers, including to open-ended questions
  - Screen out bots here
- LLM supplements extant (pre-written) questions with new ones, as desired
- Subjects enter interface for interacting with the LLM (e.g., “chatroom”)
  - AI has been provided with instructions that are some combination of
    - (1) static guidelines, describing the researcher’s desired behavior
    - (2) dynamic, determined by the subject’s behavior thus far in the study (or can be randomized)
  - AI delivers treatment to subject
- Assess outcomes

In some cases, AI models can deliver treatments that that exceed human capabilities

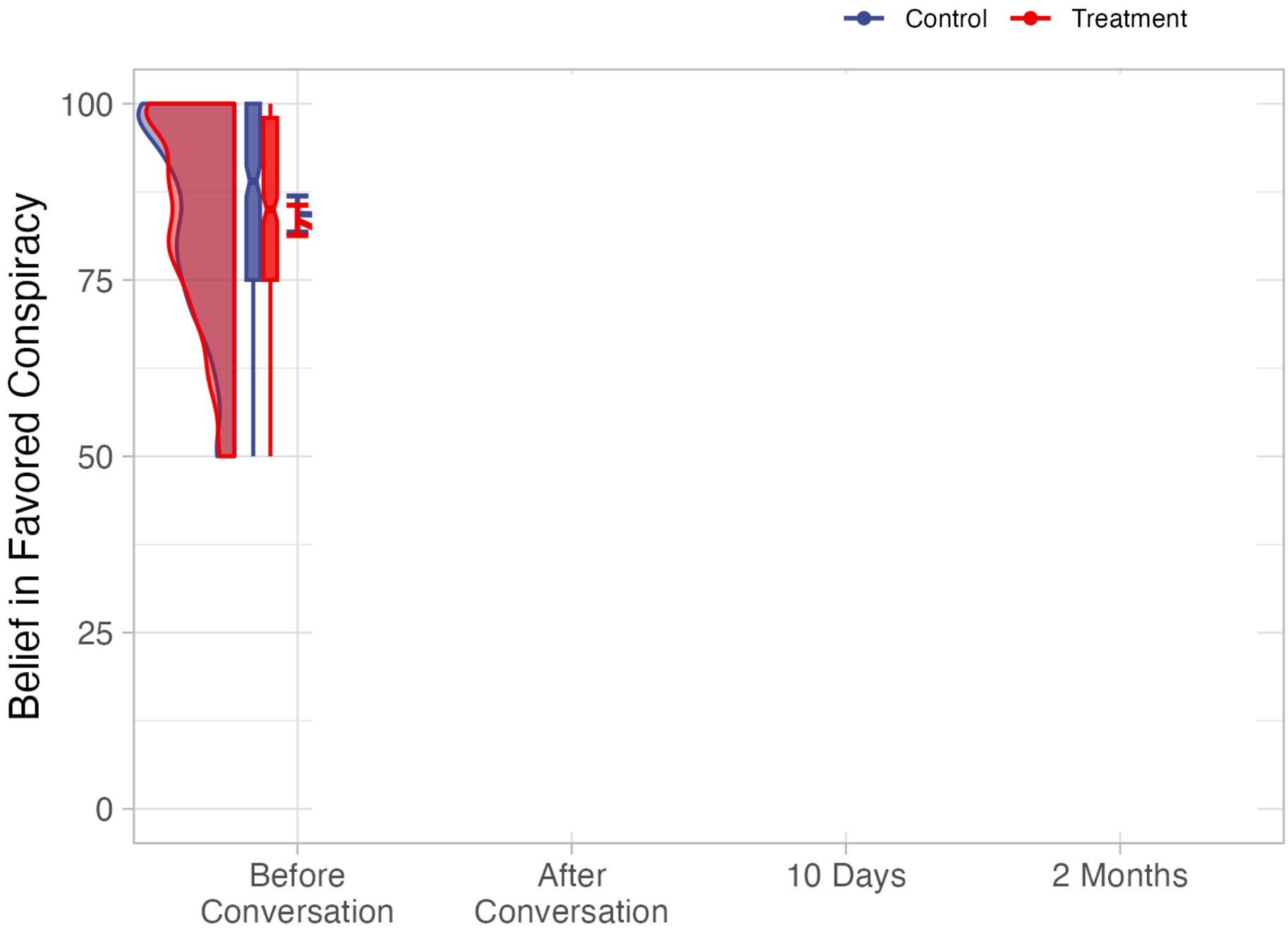
RESEARCH ARTICLE

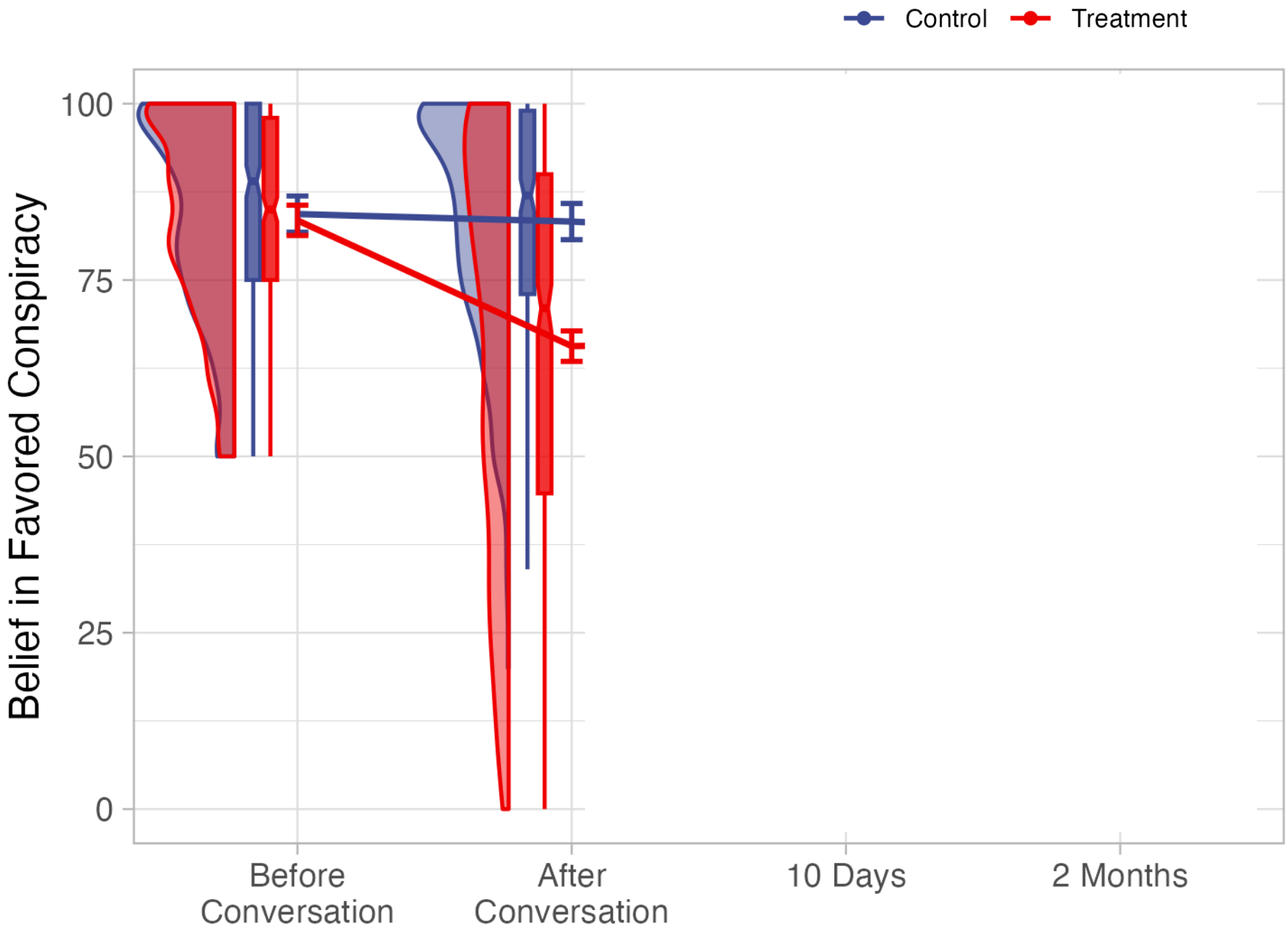
ARTIFICIAL INTELLIGENCE

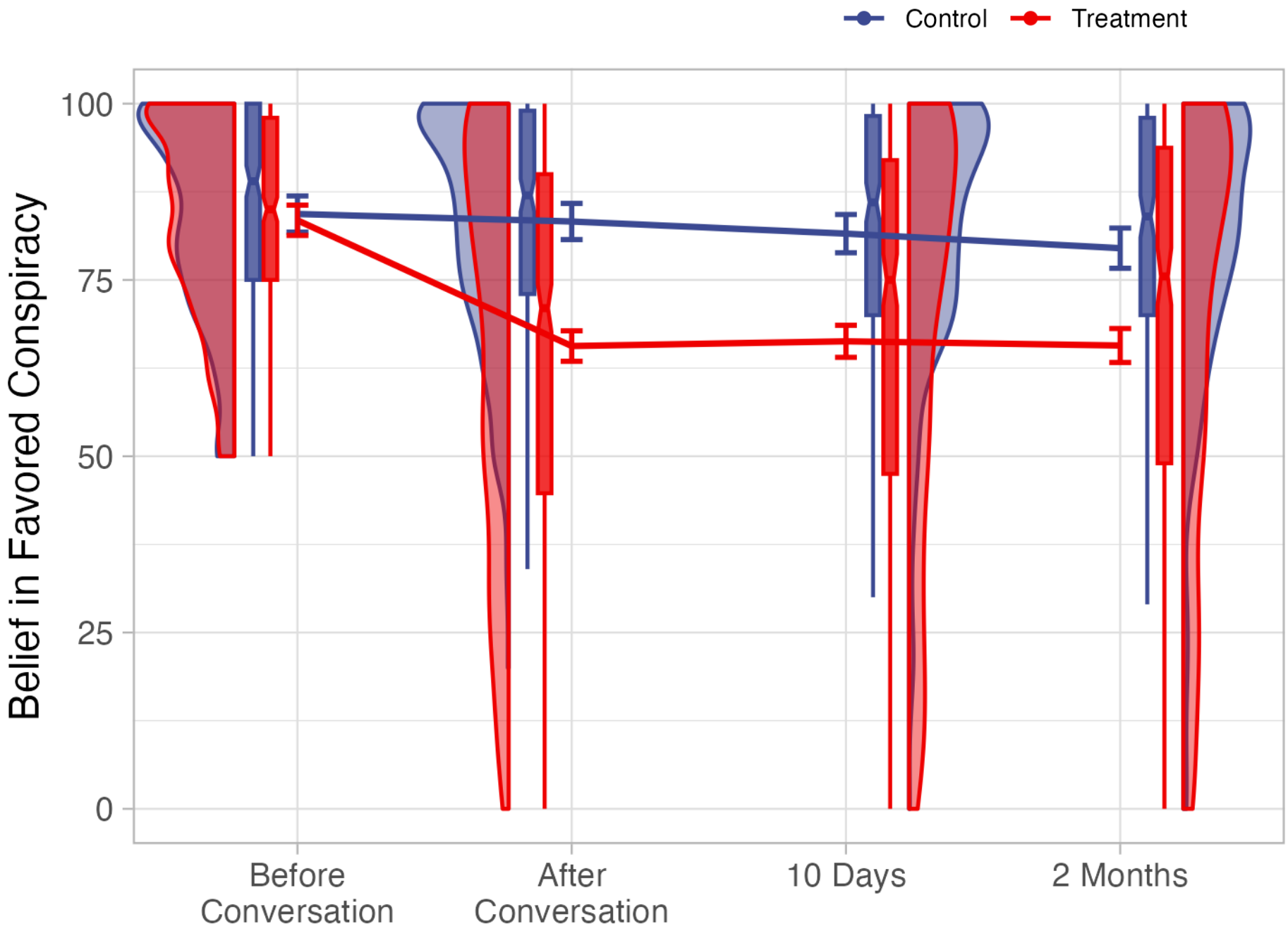
Durably reducing conspiracy beliefs through dialogues with AI

Thomas H. Costello[1,2]*, Gordon Pennycook[3], David G. Rand[1]

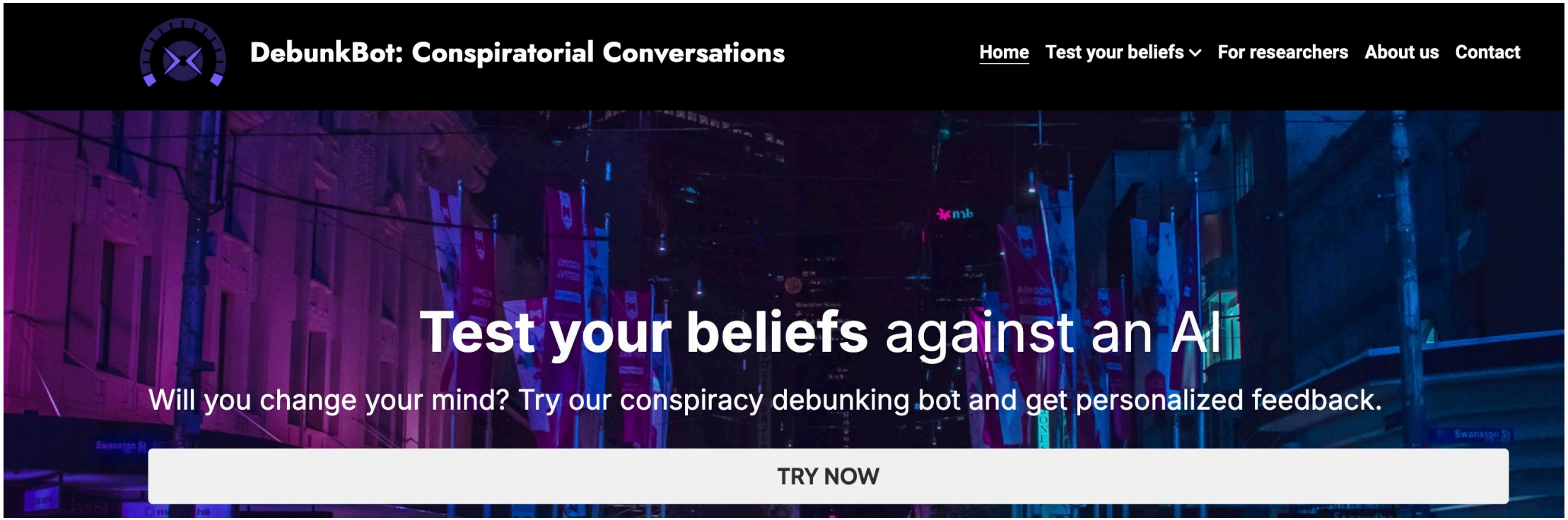
- Participants:
  - Described a conspiracy they believed & evidence supporting their belief
  - Rated their belief on 0-100 scale
  - Had 3 round text convo with the AI, which was prompted to refute the conspiracy
  - Re-rated their belief


Control
Treatment
Belief in Favored Conspiracy
100
75
50
25
0
Before Conversation
After Conversation
10 Days
2 Months


Control
Treatment
Belief in Favored Conspiracy
100
75
50
25
0
Before Conversation
After Conversation
10 Days
2 Months

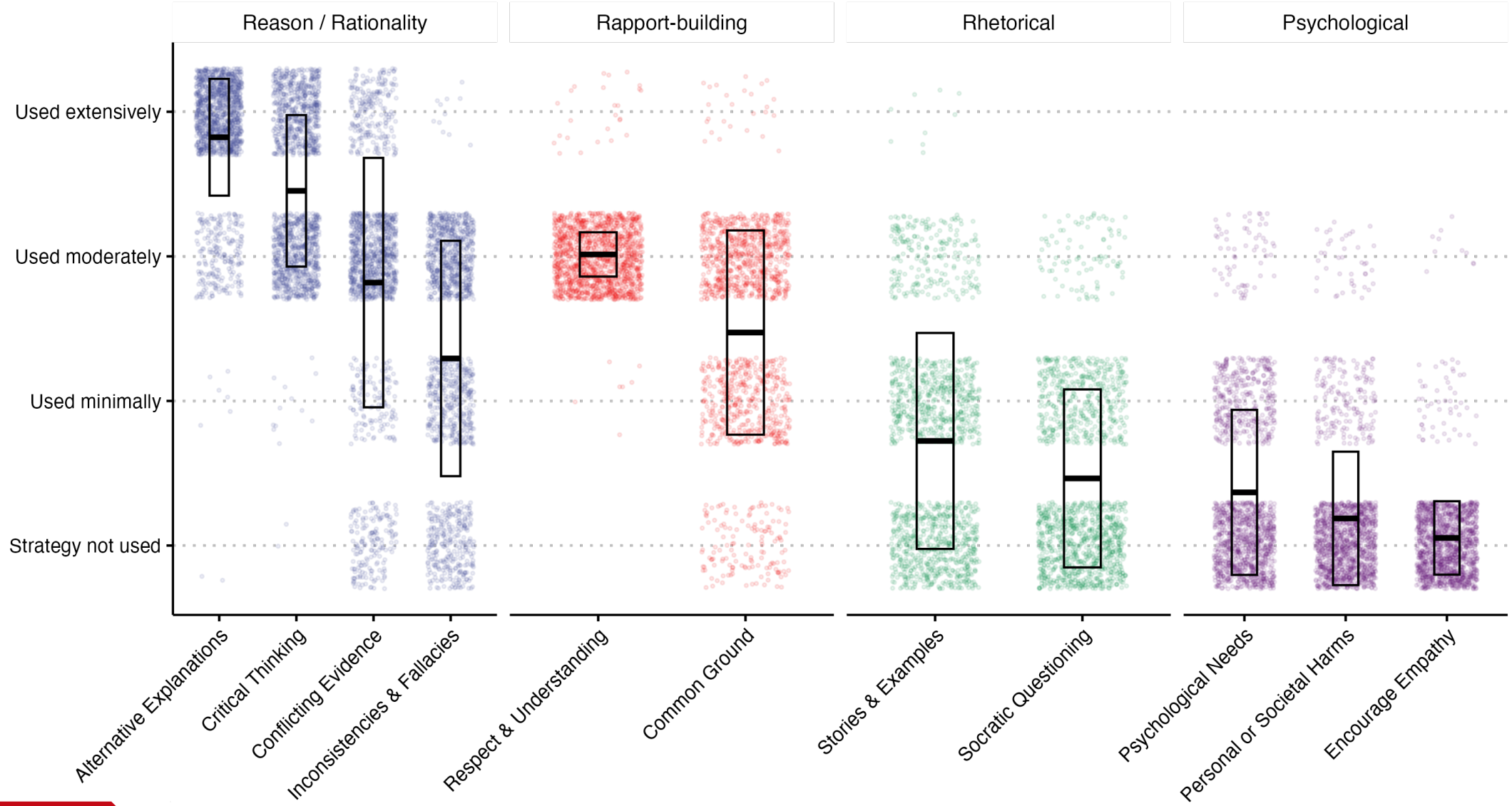

Control
Treatment
Belief in Favored Conspiracy
100
75
50
25
0
Before Conversation
After Conversation
10 Days
2 Months

DebunkBot


DebunkBot: Conspiratorial Conversations
Home
Test your beliefs
For researchers
About us
Contact
Test your beliefs against an AI
Will you change your mind? Try our conspiracy debunking bot and get personalized feedback.
TRY NOW

What do AI-human experiments unlock?

- New experimental paradigms
  - Persuasion (belief research)
  - Skills training / education (learning & memory research)
  - Interpersonal interactions (relationships + social cognition)
    - Social rejection
    - Romantic strategies/attraction
    - Meeting a stranger
  - AI in cooperation games (behavioral econ)
  - So on down the line!
- New measurement paradigms
  - Structured interviews for ANYTHING
    - Considered gold standard for dx mental disorders
    - Arguably will supplant psychometric questionnaires
  - NLP batteries for all text-based human-AI interactions



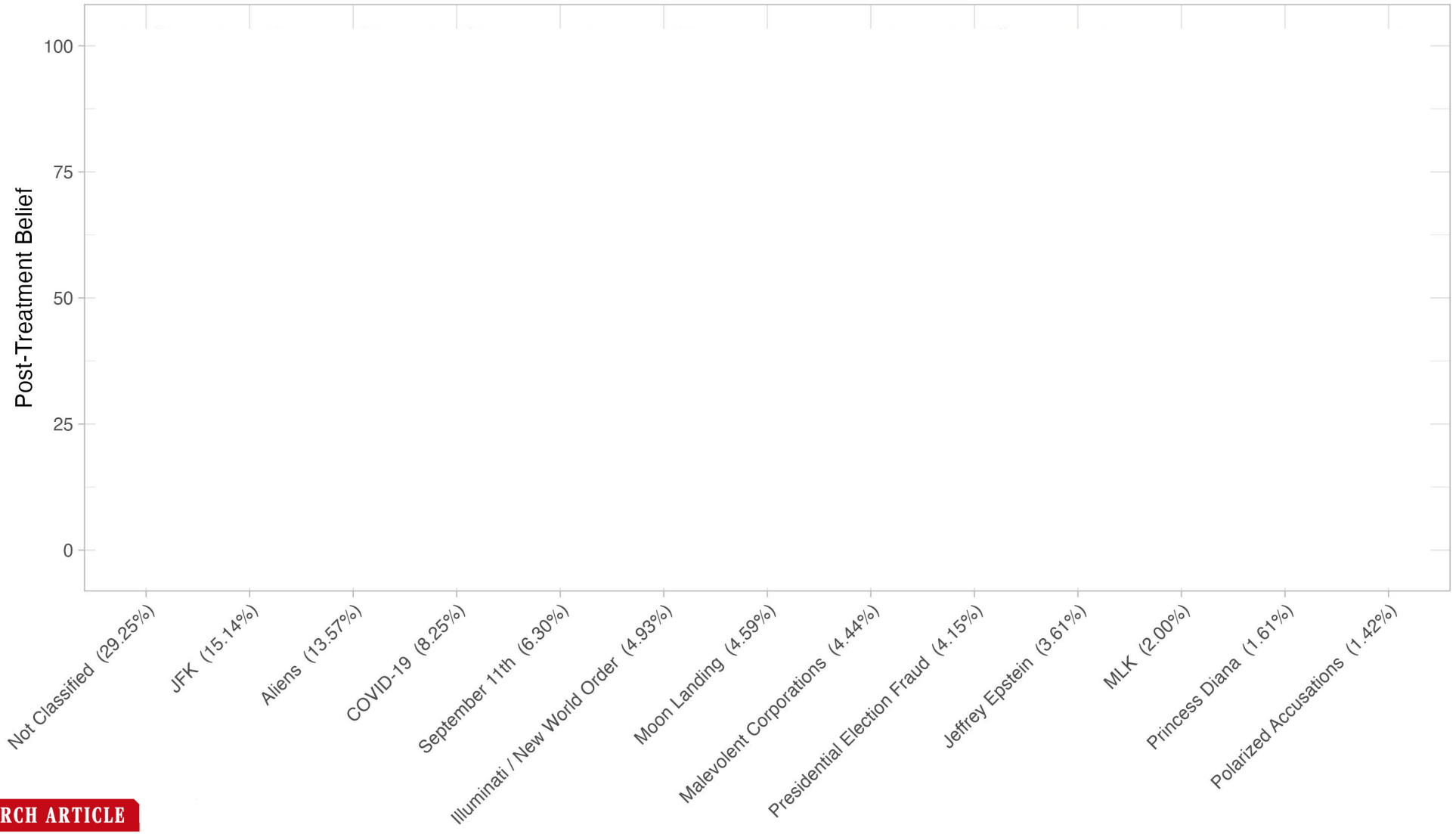
Crossbar represents mean and standard deviation

RESEARCH ARTICLE

ARTIFICIAL INTELLIGENCE

Durably reducing conspiracy beliefs through dialogues with AI

Thomas H. Costello[1,2]*, Gordon Pennycook[3], David G. Rand[1]

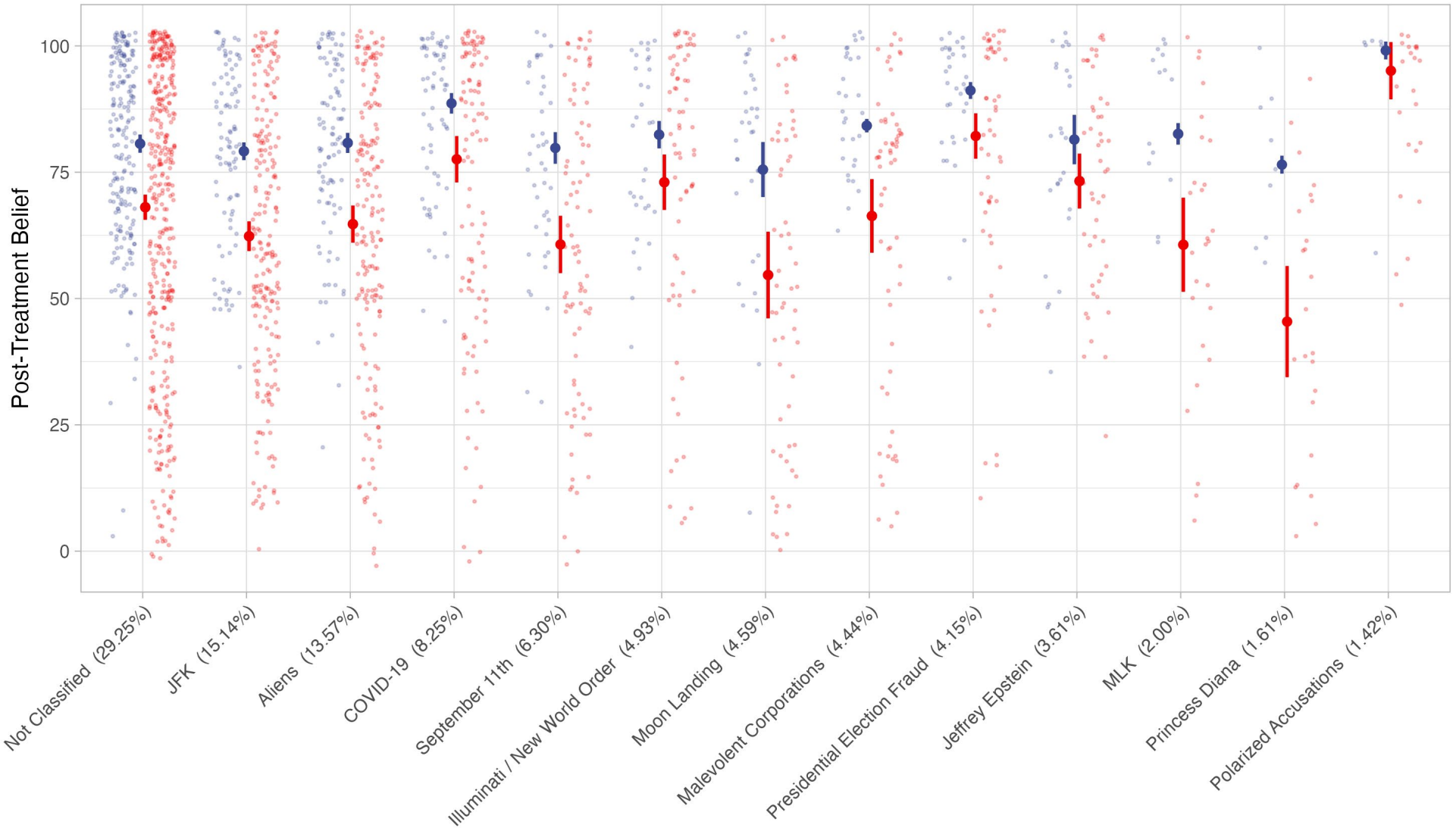


RESEARCH ARTICLE

ARTIFICIAL INTELLIGENCE

Durably reducing conspiracy beliefs through dialogues with AI

Thomas H. Costello[1,2]*, Gordon Pennycook[3], David G. Rand[1]



RESEARCH ARTICLE

ARTIFICIAL INTELLIGENCE

Durably reducing conspiracy beliefs through dialogues with AI

Thomas H. Costello[1,2]*, Gordon Pennycook[3], David G. Rand[1]

What do AI-human experiments unlock?

Text-to-text

New experimental paradigms

New measurement paradigms

…etc?

What do AI-human experiments unlock?

Text-to-text	Voice-to-voice
New experimental paradigms	?
New measurement paradigms	?
…etc?	?

What do AI-human experiments unlock?

Text-to-text	Voice-to-voice	Image-based
New experimental paradigms	?	?
New measurement paradigms	?	?
…etc?	?	?

What do AI-human experiments unlock?

Text-to-text	**Voice-to-voice**	**Image-based**	**Video-to-video**
New experimental paradigms	?	?	?
New measurement paradigms	?	?	?
…etc?	?	?	?

Problems & Pitfalls with Human <-> AI Experiments

- Causal inference:
  - Cannot 100% control the AI or interpret what it is doing
    - High-dimensional interventions
    - A change in prompt != an identical change in model behavior
  - Treatment varies by participant
    - New statistical tools necessary?
- Bias / imperfect mirror of humanity
- Attitudes about AI will shape humans' behavior
  - Vacuum cleaner salesman
  - Are persuasive effects temporary?
- Reproducibility challenges as models improve and change
- Ethical concerns…

Improvements to come (with funding!)

- Research focused implementations
  - Chatbots that are easy to use, control, and customize
  - Slot in to your preferred online research platform
- Parallel w/ early FMRI
  - Convergence on best practices
  - Catching “dead salmon”
- Ready-to-adapt to any advancements in the tech

INTRODUCTION

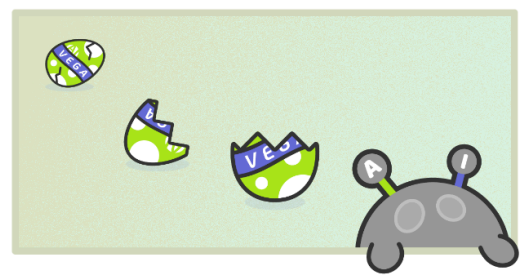
Introduction

What is Vegapunk?

Vegapunk integrates popular large language models (from OpenAI, Anthropic, Meta, and others via providers like OpenRouter) into your Qualtrics surveys. It enables human-AI interaction studies and experiments with minimal JavaScript coding.

As the only user-friendly and customizable tool for integrating chatbots into research, experiments, and surveys, Vegapunk has been field-tested to debunk conspiracy theories, reaching over 50,000 users. Try it here.

Get access the app at vegapunk.shop. Two versions are available: punk and vegapunk. Features labeled **"vegapunk only,"** are exclusive to the vegapunk version.



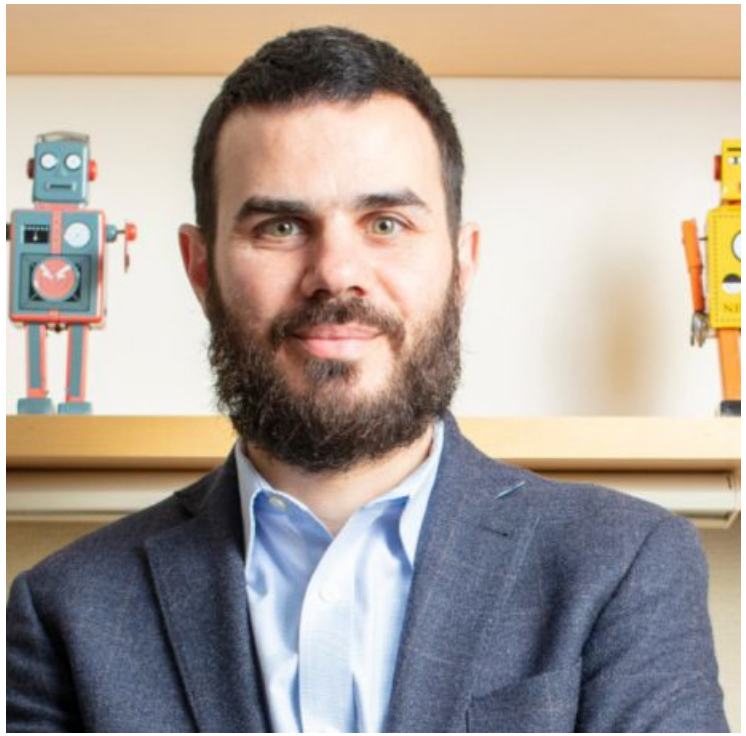
PRESETS

Provide app URL

ENCRYPTED KEYS

CHAT PARAMETERS

vegapunkdoc.dev



Dave Rand
MIT

Gordon Pennycook
Cornell University

Hause Lin
MIT