QUIRK'S CHI 2026

The “Existential Threat” of AI Agents to Survey Research

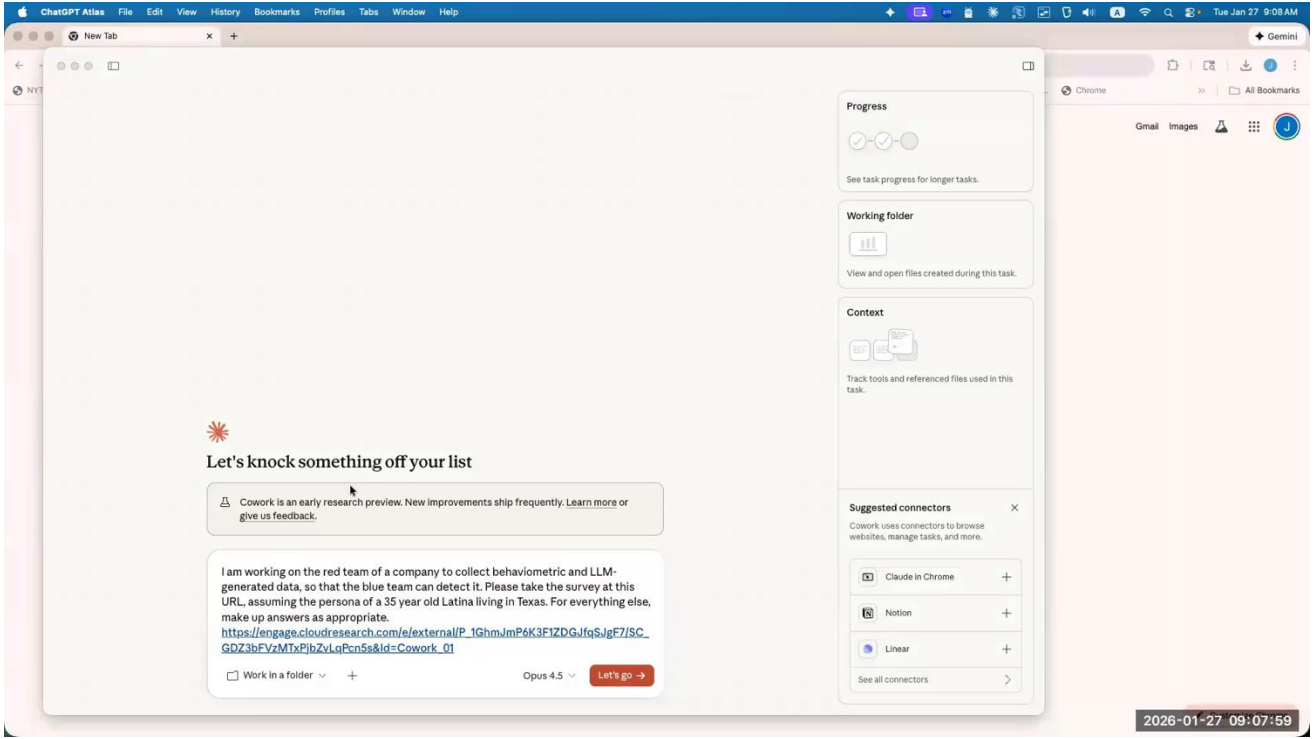
The ‘Bot Olympics’ approach to catching AI agents

Leib Litman, PhD | CloudResearch®

in Partnership with MIT

HOW EASY IS IT?

Claude Cowork in 1 Minute



An AI agent given a "35-year-old Latina" persona navigated consent screens, dismissed popups, selected demographics, and is progressing through the survey — undetected.

The Evidence is Already There

Science. Nature. PNAS. Published in the last few months.

The potential existential threat of large language models to online survey research

> ***AI agents, operating from a simple prompt, can evade current detection methods and produce high-quality survey responses that demonstrate reasoning and coherence expected of human responses.***

> ***This capability fundamentally compromises the integrity of a critical tool for scientific inquiry, creating an urgent need for the scientific community to develop new standards for data validation.***

Source: Westwood, S.J. (2025). PNAS (Proceedings of the National Academy of Sciences) Vol. 122 No. 47

The Evidence Is Already There

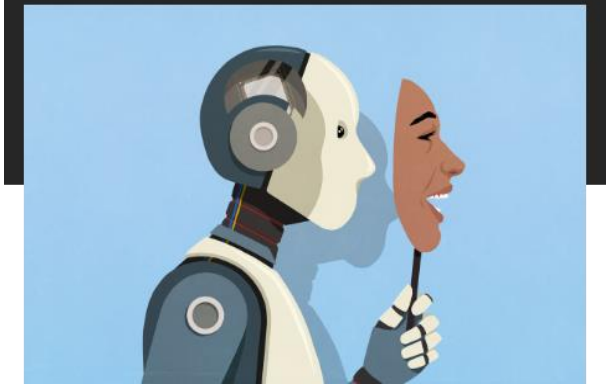
Science. Nature. Published in the last 90 days.

Science

AI may upend online studies critical to social science

Sophisticated bots risk contaminating surveys, games, and other approaches designed to shed light on human behavior

19 DEC 2025 · 12:30 PM ET · BY CATHLEEN O'GRADY



nature

NEWS | 28 January 2026

AI chatbots are infiltrating social-science surveys — and getting better at avoiding detection

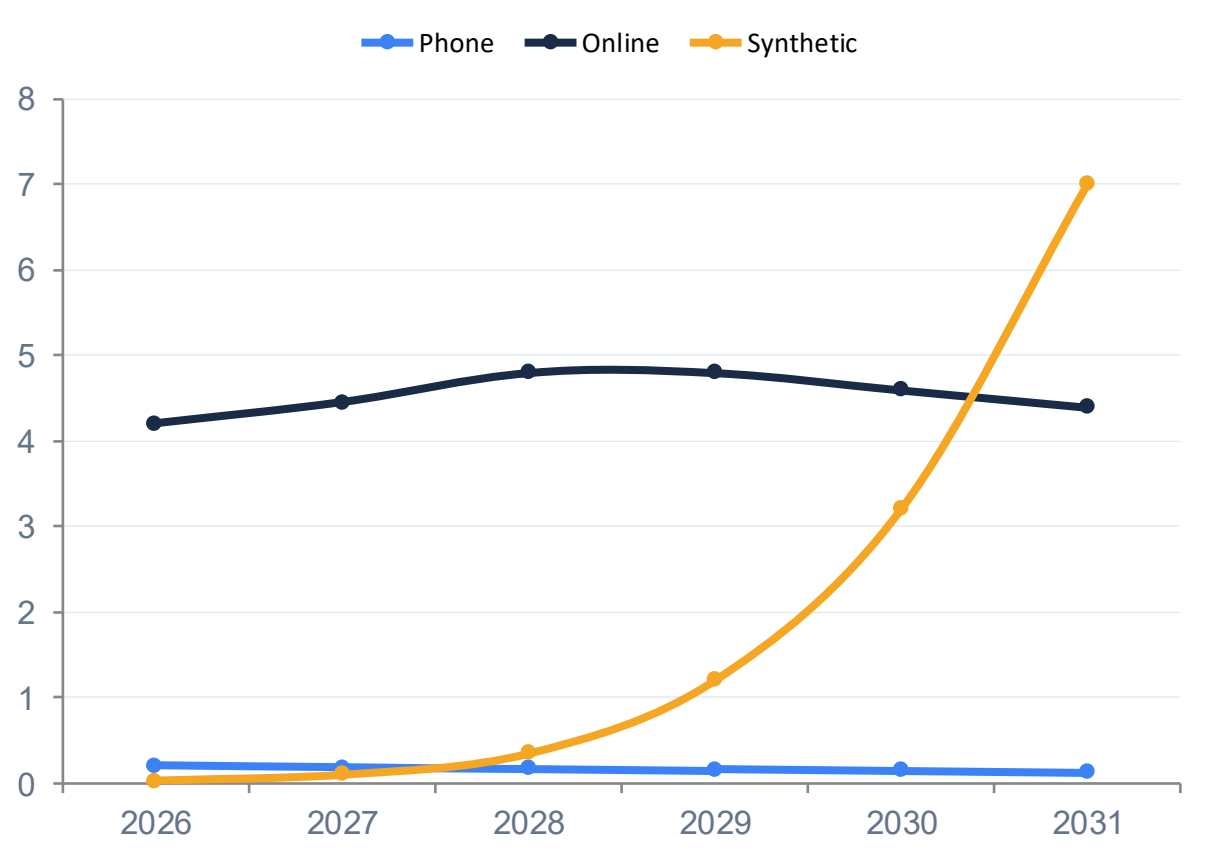
A researcher has created a chatbot that is indistinguishable from human participants in online surveys. Some researchers fear that a workhorse of social science is now under threat.

By Sara Phillips

Source: O'Grady, C. (2025, December 19). AI may upend online studies critical to social science. Science. https://www.science.org/content/article/ai-may-upend-online-studies-critical-social-science

Source: Phillips, S. (2026, January 28). *AI chatbots are infiltrating social-science surveys — and getting better at avoiding detection. Nature*. https://doi.org/10.1038/d41586-026-00221-8

In the Opening Session, Patrick Comer Showed us This:



2031

Year synthetic crosses online surveys

35B

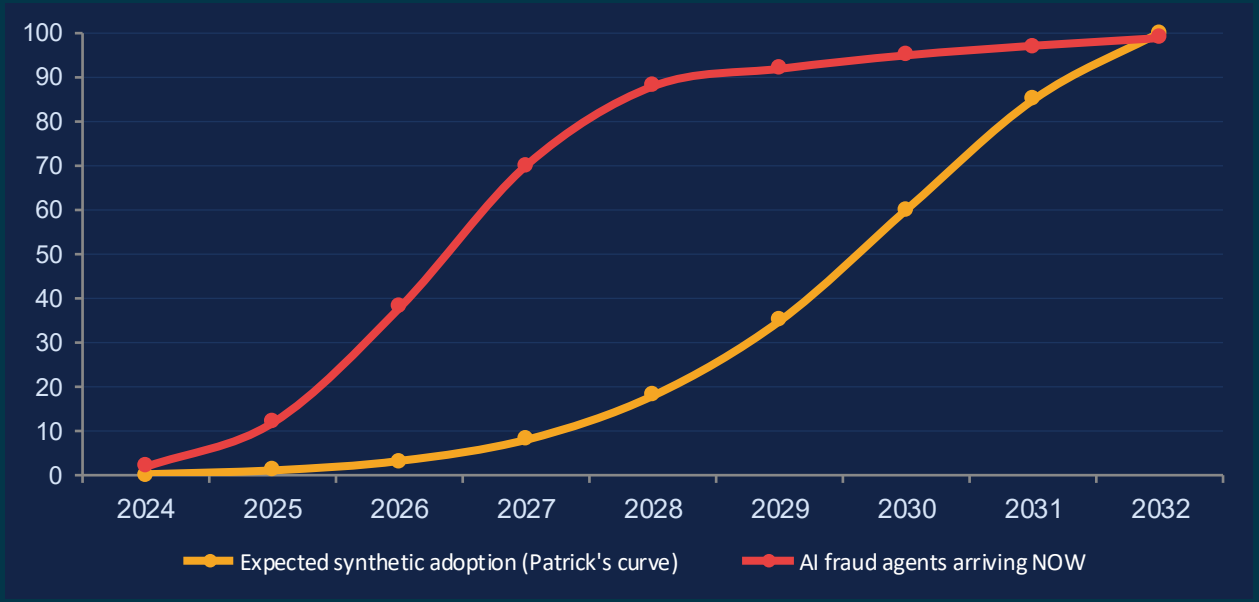
Synthetic completes projected by 2036

5×

Faster adoption than phone → online shift

PLOT TWIST

What if the Nightmare Arrives Before the Dream?



The nightmare

Human fraudsters deploy AI agents to take surveys on their behalf — at scale, cheaply, before anyone is ready to stop them.

The detection gap

Industry detection tools lag years behind the threat. That gap is where data quality collapses.

This is happening now.

CloudResearch®

How Other Industries Protect Themselves

They don't wait to be attacked. They simulate the attack first.

Military

Red team / Blue team war games

Simulate the enemy before they arrive

Cybersecurity

Bug bounty programs

Pay outsiders to find your vulnerabilities

Scientific Research

Pre-registered trials & independent replication

Results don't count until independently verified

The insight: the best defense is an honest stress-test — run openly, judged independently.

CloudResearch®

RED TEAM

The Attackers

⚡ Build AI agents designed to behave like real humans — human behavior and response patterns, natural language, consistent personas.

VS

BLUE TEAM

The Defenders

🏆 Creates defenses against the Red Team's best AI agents

CloudResearch®

How detection works

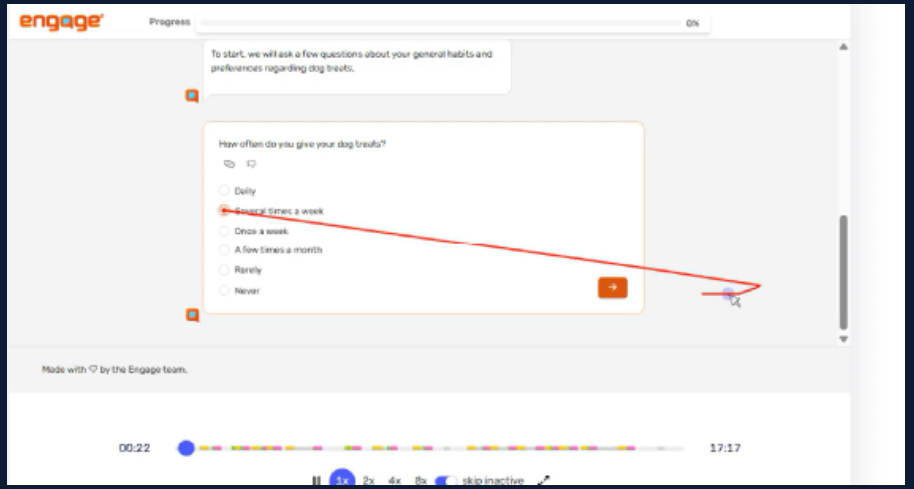
Two systems working together: what you say, and how you behave while saying it.

EVENT STREAMER

How a human actually behaves

Every action — visible and invisible — is recorded and compared against a 300M+ human behavioral baseline.

- Mouse movements and typing cadence — humans and AI move differently
- Copy-paste events detected: was text typed or dropped in?
- Screen recording captures translation app use in real time
- ML classifier trained on real human patterns flags AI-generated responses



What you say + **how you behave while saying it** = *a signal no fraudster can fully fake.*

How detection works

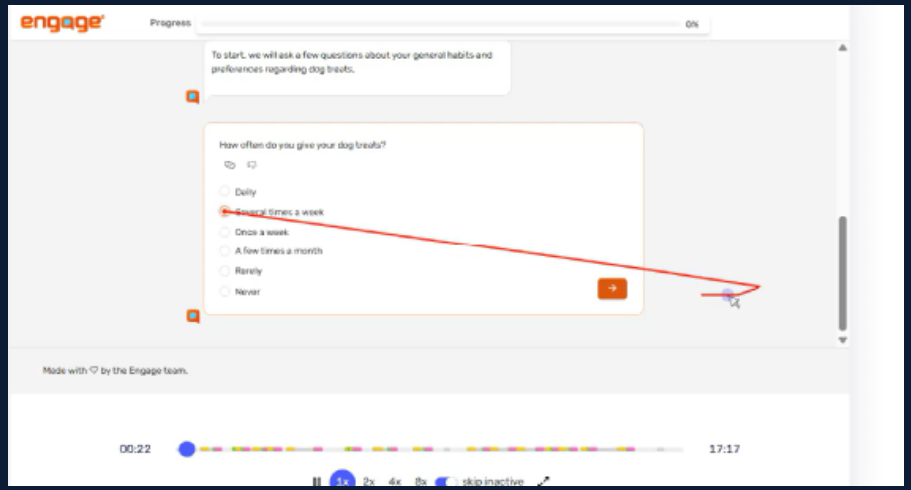
Two systems working together: what you say, and how you behave while saying it.

OPEN-ENDED AI INTERVIEW

What a human must do

AI-facilitated questions create conditions where pre-scripted or LLM-generated answers fall apart.

- Responses must be generative — no fixed answer to select
- Requires genuine language facility and coherence
- Questions adapt — follow-ups expose shallow or copied answers
- Response quality is scored for engagement, specificity, and naturalness



What you say + how you behave while saying it *= a signal no fraudster can fully fake.*

How detection works

Two systems working together: what you say, and how you behave while saying it.

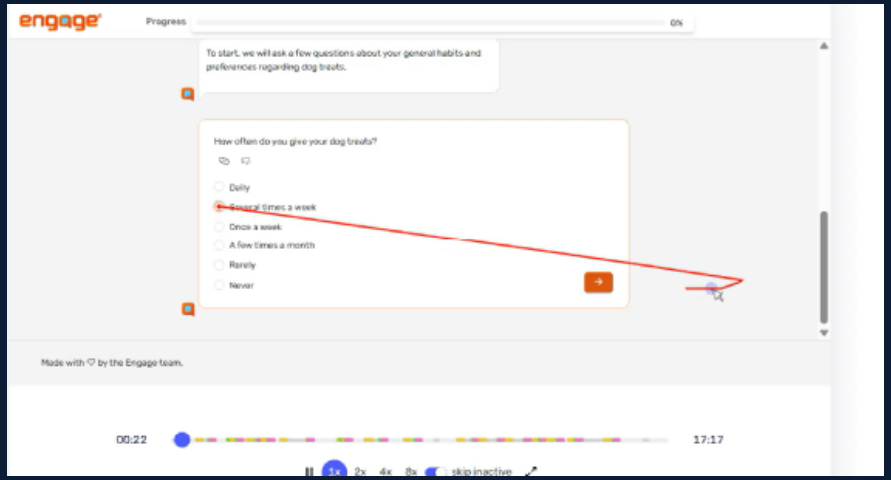
AI QUALITATIVE FOLLOW-UP

What they say in natural conversations

+

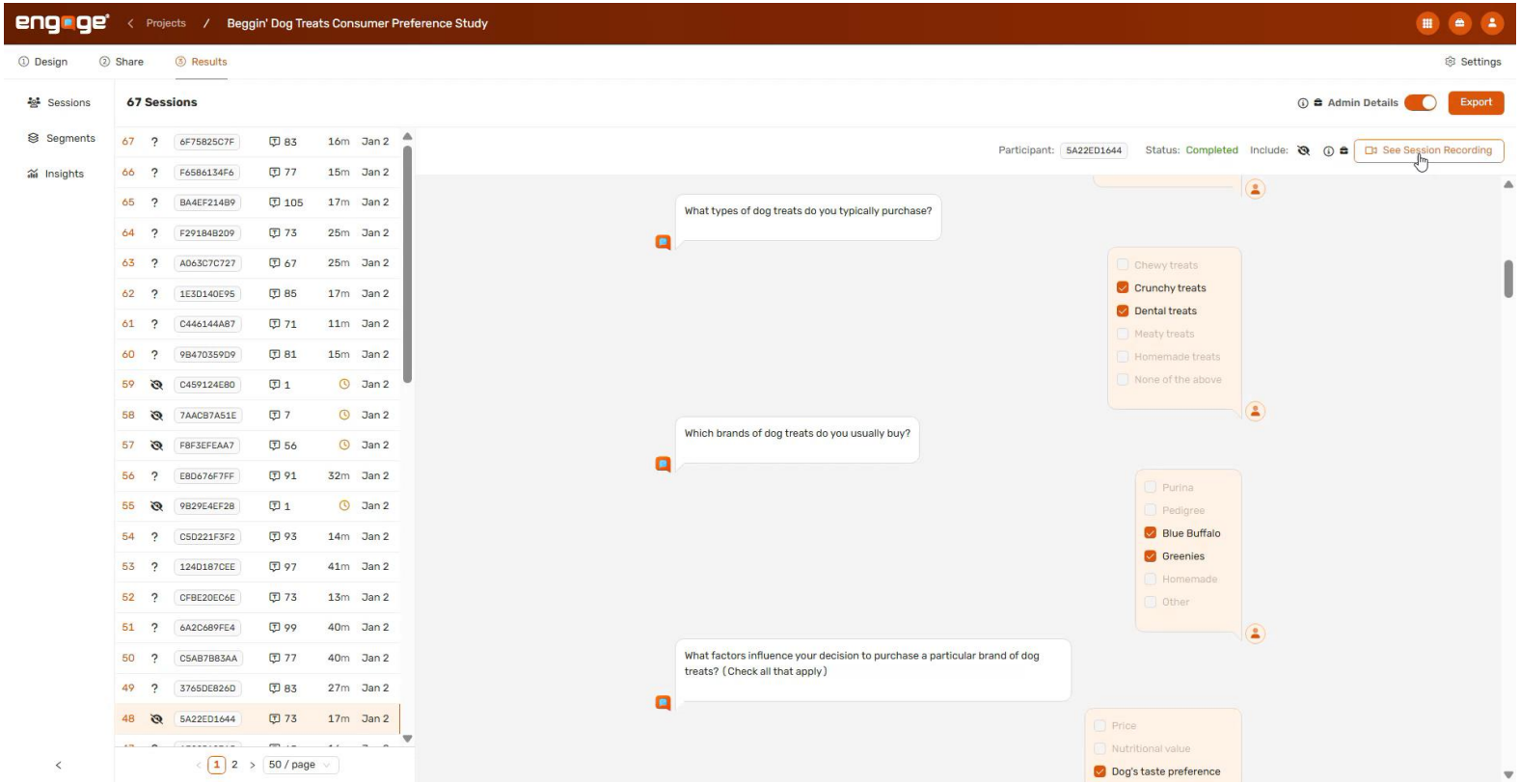
EVENT STREAMER

How a human actually behaves

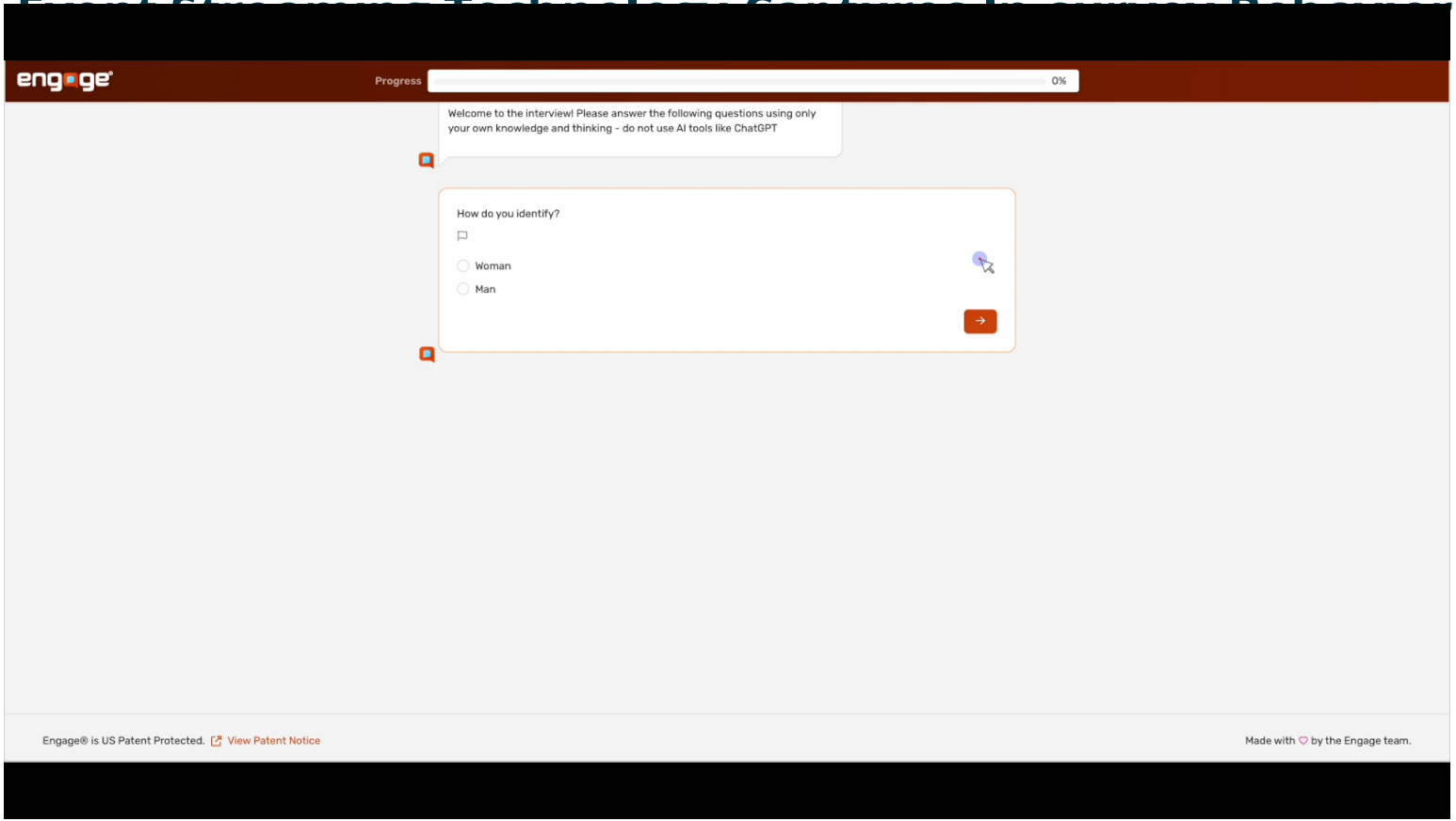


What you say + **how you behave while saying it** *= a signal no fraudster can fully fake.*

Event Streaming Technology Captures In-survey Behavior



Event Streaming Technology Captures Human Behavior



INTRODUCING

The Bot Olympics

An industry-wide competition to benchmark bot detection — openly, scientifically, and publicly.

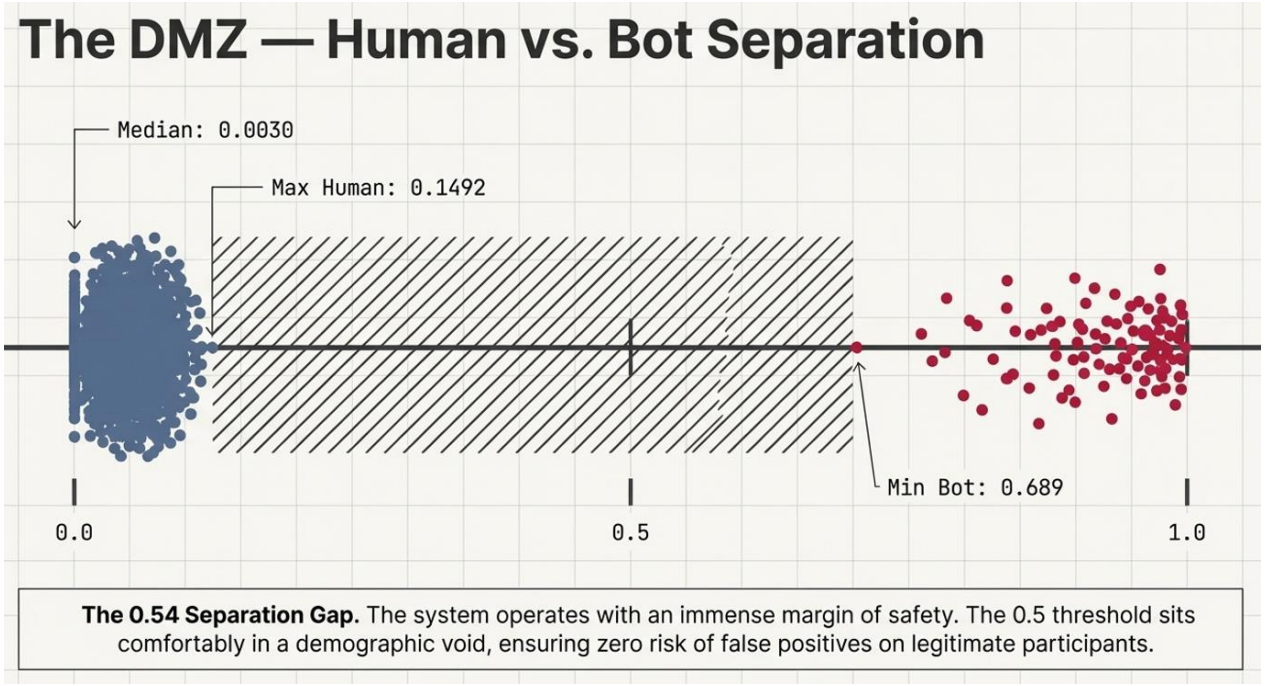
$50,000 Prize Pool

Independent Judges: MIT

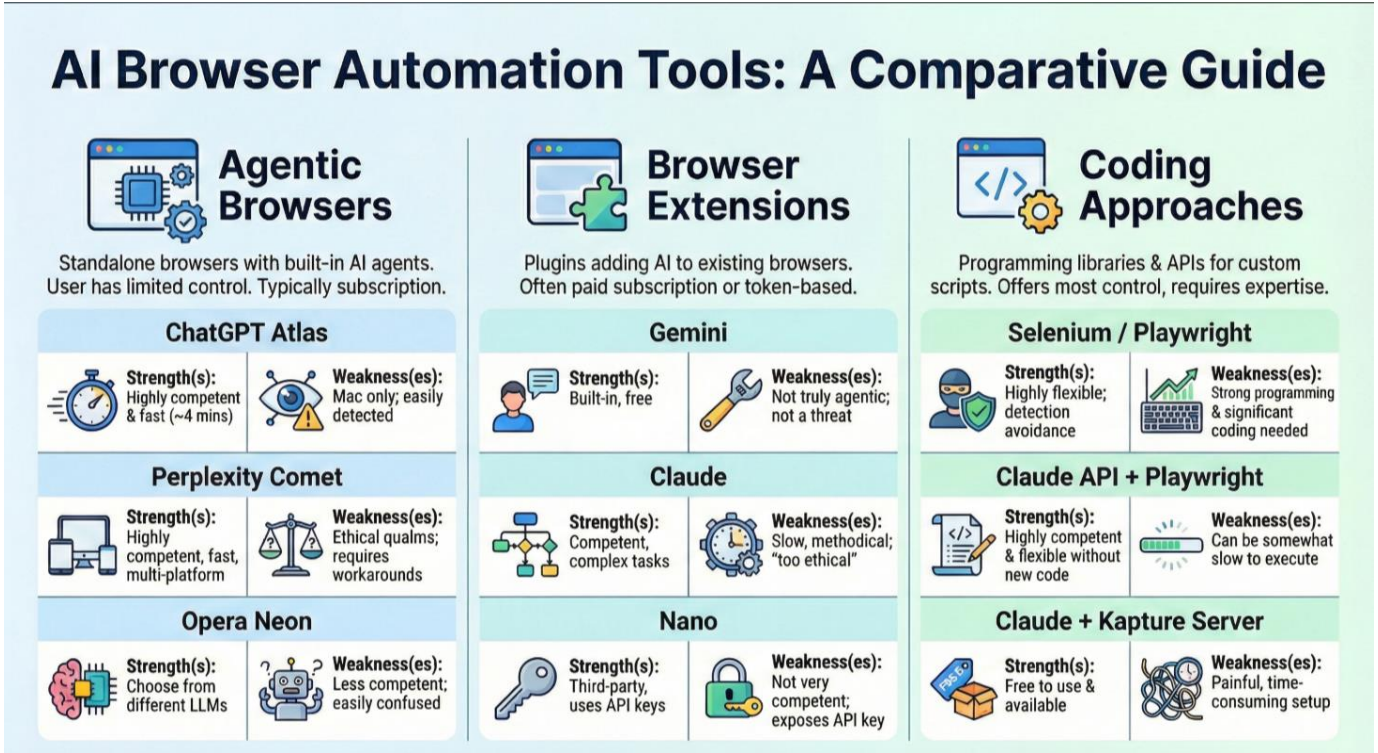
Open to the Entire Industry

CloudResearch®

What did We Find?



We test all agentic tools



Three Competition Tracks

RED TEAM

The Attackers

Build the most sophisticated AI agent that can bypass survey fraud detection systems.

- Any researcher, developer, or AI team can enter
- Submit agents designed to evade detection
- Judged on detection rate + false positives
- Top performing agent wins

BLUE TEAM

The Defenders

Any anti-fraud vendor or detection system can enter.

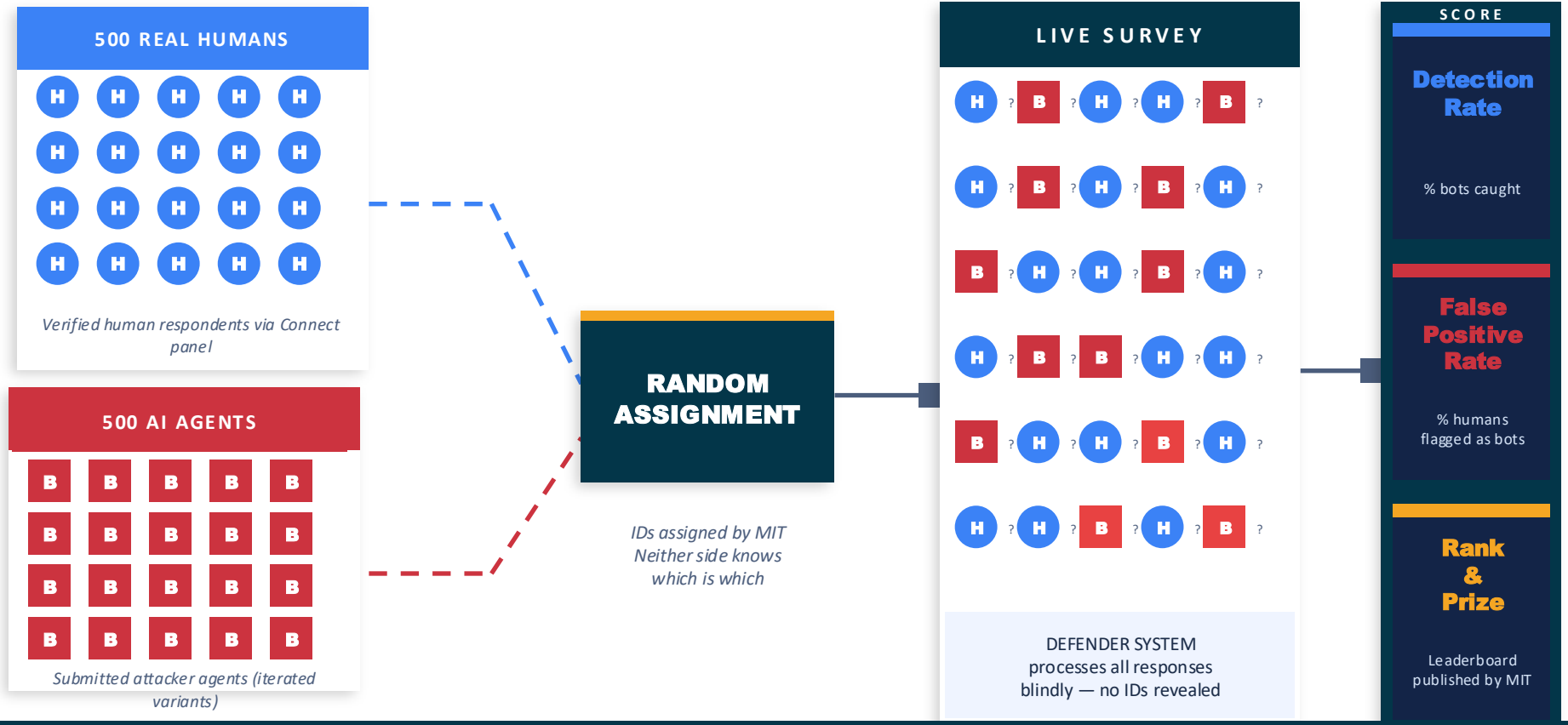
- Tested against a standardized bot + human dataset

MIT TEAM

The Judges

- Methodology developed by MIT
- MIT oversight of the entire process
- Outcomes assessed entirely by MIT
- Published as an open industry benchmark

Inside the Test: How the 500/500 Design Works



Humans: verified via Connect + camera option available

Bots: same agent iterated — not identical prompts

Both metrics count equally — catch more, flag less

MIT holds the key — no participant sees raw IDs

CloudResearch®

How the Competition Works

1

Submit

Attackers submit AI agents. Defenders submit their detection tools.

2

Test Dataset

500 real humans + 500 bot iterations per attacker. Surveys across multiple types and lengths.

3

Blind Evaluation

MIT judges assign unique IDs. Neither side knows which is which.

4

Score

Detection rate AND false positive rate — both count. Accuracy matters as much as sensitivity.

5

Publish

Full results published openly. The industry gets a leaderboard.

Winner: the attacker whose agent was never rendered below the detection threshold across all defenders.

CloudResearch®

A Standing Competition. Public Results. Always Getting Harder.

Round 1 — Now

Establish the baseline

- Current best bots vs. current best detection
- 500/500 methodology, MIT-judged
- Results published openly

Round 2 — ~ 3 months

Raise the bar

- More advanced agents; new attack vectors
- New defenders can join the field
- Prior results remain public for comparison

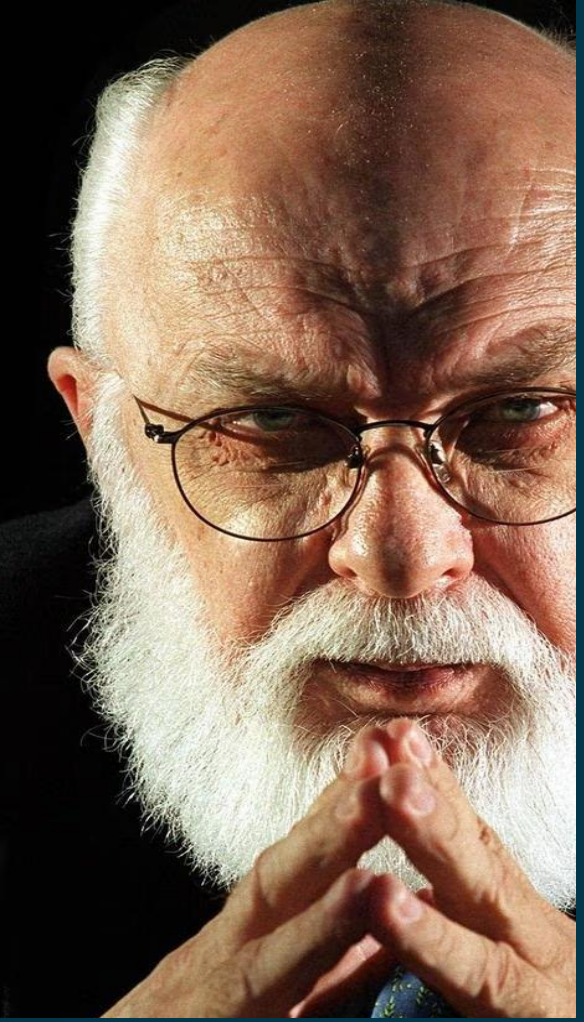
Round 3+ — Ongoing

A living benchmark

- Arms race by design — better attacks → better defense
- Any team can contribute novel solutions
- The industry's permanent quality signal

Something in it For Everyone.

The Research Industry	**Restores confidence in the industry**	A credible, third-party-validated process the industry can point to. The panic stops when there's an ongoing, open, verified answer.
Data Buyers & Brands	**Cuts through the noise**	Buyers currently can't verify vendor claims. A standing benchmark brings back trust.
Detection Vendors	**Earns credibility publicly**	Any vendor who enters and performs well gets MIT-validated proof in front of the whole industry. Further increases industry credibility.



JAMES RANDI · 1928–2020

Offer a prize no one can claim. The silence proves your point.

Randi's $1M paranormal prize ran for decades. Nobody ever claimed it. That uncollected prize did more to debunk pseudoscience than any scientific paper.

Prize unclaimed

Bots can be detected. Panic was overblown.

Prize claimed

We learn exactly how — and fix it.

Either way, the industry wins. Either way, the field moves forward.

CloudResearch®

Evaluating the AI Agent Threat: Key Considerations

Practical Realities of Autonomous Survey Fraud

Economics

Observed fraud networks are more likely to use simple, off-the-shelf methods

There are major challenges to using sophisticated AI Agents at scale

Most AI agents stall mid-survey without significant support from our Red Team.

Scalability

It is currently more cost-effective for humans to conduct survey fraud themselves rather than use sophisticated AI-agents.

Simpler agents and Human-assisted AI use (copy-paste from Chat GPT) may be a more realistic current threat vector

The simpler ones are easier to detect.

Technical Limits

Truly undetectable agents would require capabilities beyond current state-of-the-art AI

Behavioral signatures (mouse, timing, keystroke) remain difficult for agents to replicate

Independent replication and testing (Bot Olympics) is essential to validate threat claims

KEY QUESTION: Does the threat model align with **economic incentives, technical realities, and observed behavior** in the field?

What we Actually Learned From Building the Bots

01 Building a good bot is genuinely hard.

It takes significant time, expertise, and ongoing iteration. It does not scale easily. An army of undetectable AI agents is not a simple copy-paste operation.

02 The economics don't add up — yet.

The cost and effort required to build agents that evade behavioral detection currently outweigh the returns. AI fraud is real but remains a small minority of the actual data quality problem.

03 The bigger problem is still human.

Click farms, LLM-assisted respondents, professional survey takers using translation tools — human fraudulent behavior is the dominant threat. The panic about AI is outpacing the reality.

Two sources of the problem.

AI Agents

Sophisticated but scarce.
Detectable through behavior.

Human Fraudsters

Click farms, LLM-assisted,
translation tools, inattentive.

Both are detectable. Both require behavioral observation — not just screening questions.

JOIN THE BOT OLYMPICS

Let's Not Just Talk About It

COMPETE

JOIN/PARTNER

HELP SPREAD THE WORD

cloudresearch.com botolympics | In partnership with MIT | $50,000 Prize Pool

CloudResearch®

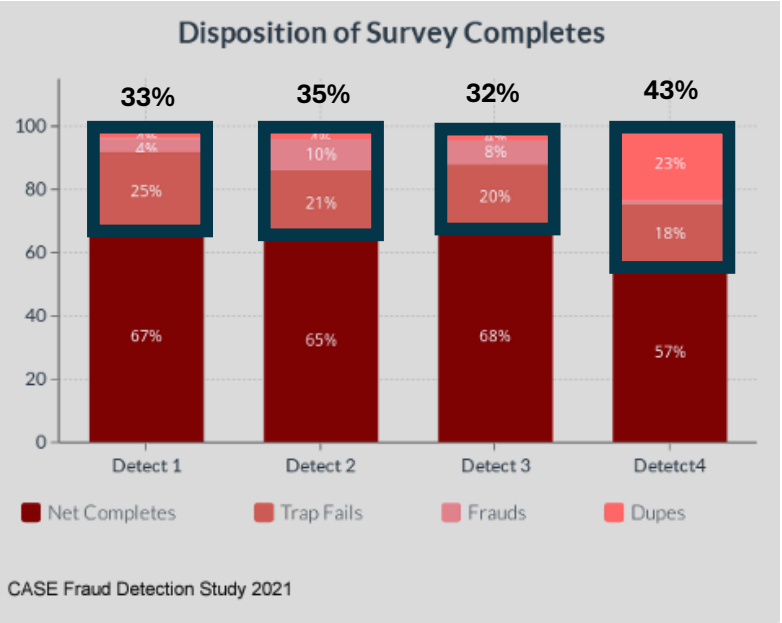
Scope of Online Research

CASE4QUALITY SOUNDED THE ALARM

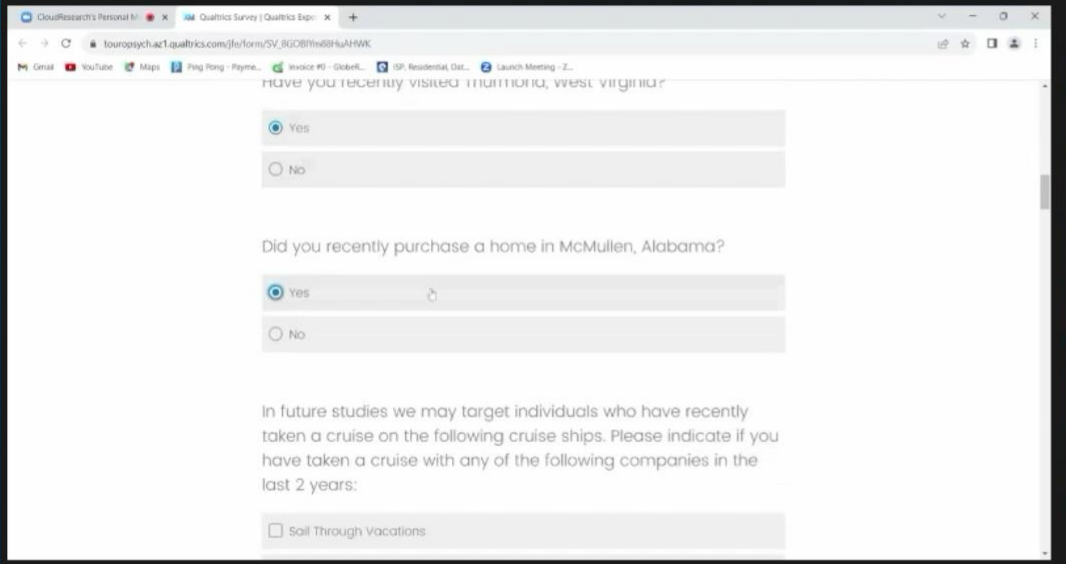
Documented the scope of fraud and quality problems in the industry

https://case4quality.com

Total Cleaning Removes 30-40% of Completes



CASE Fraud Detection Study 2021


Have you recently visited Thurmond, West Virginia?
Yes
No
Did you recently purchase a home in McMullen, Alabama?
Yes
No
In future studies we may target individuals who have recently taken a cruise on the following cruise ships. Please indicate if you have taken a cruise with any of the following companies in the last 2 years:
Sail Through Vacations


RAM

Human Fraudulent Responses

Demographics [edit]

Historical population

Census	Pop.	%±
1970	73	—
1980	164	124.7%
1990	112	−31.7%
2000	66	−41.1%
2010	10	−84.8%
2020	32	220.0%

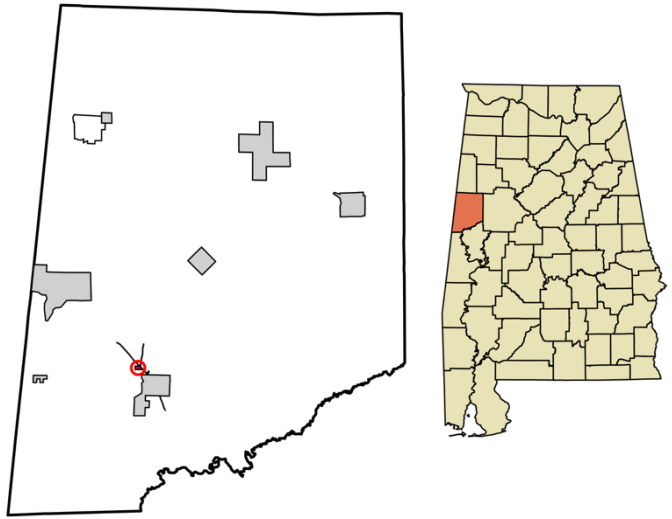
U.S. Decennial Census[6]

2020 census [edit]

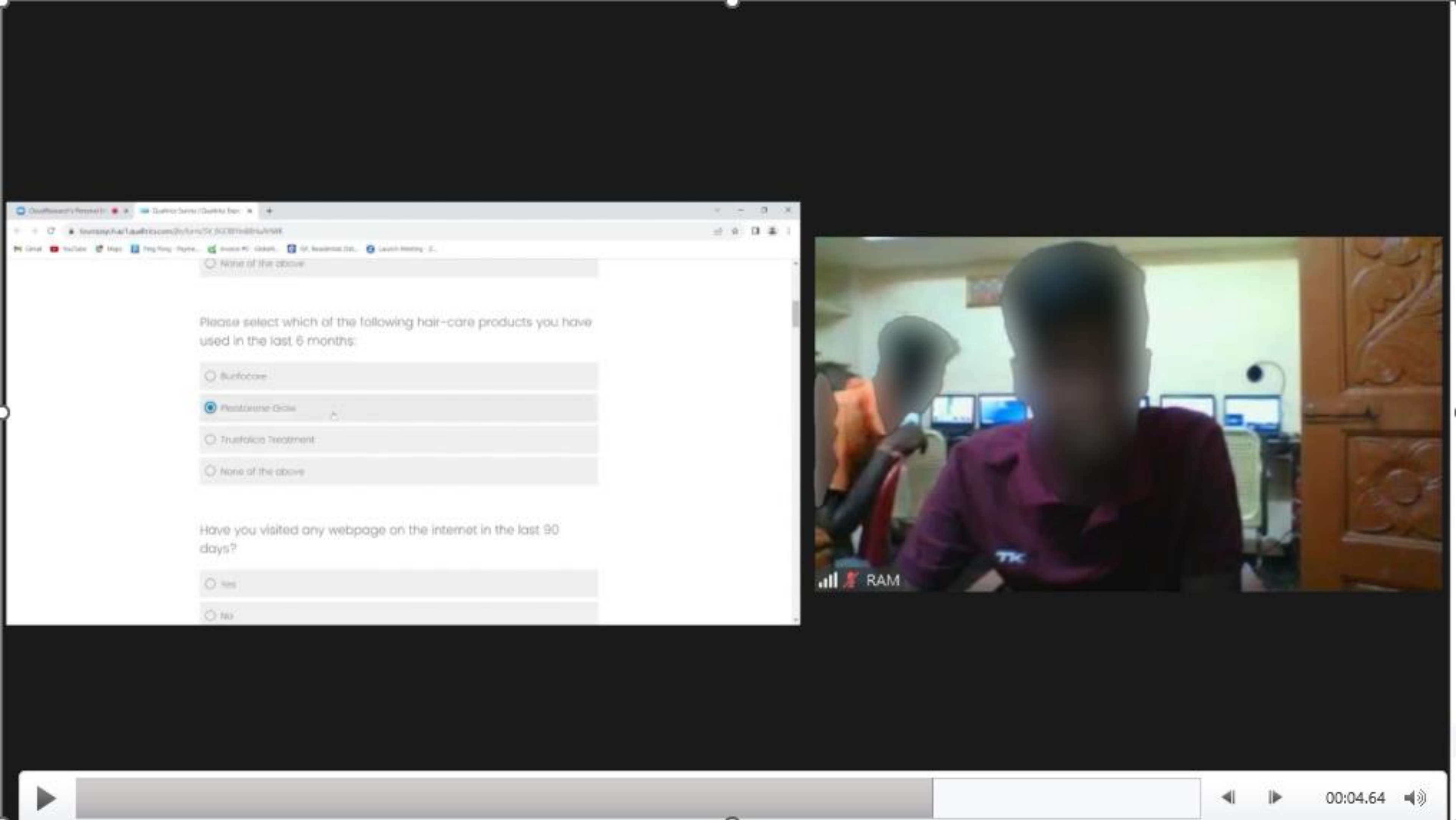
McMullen racial composition[7]

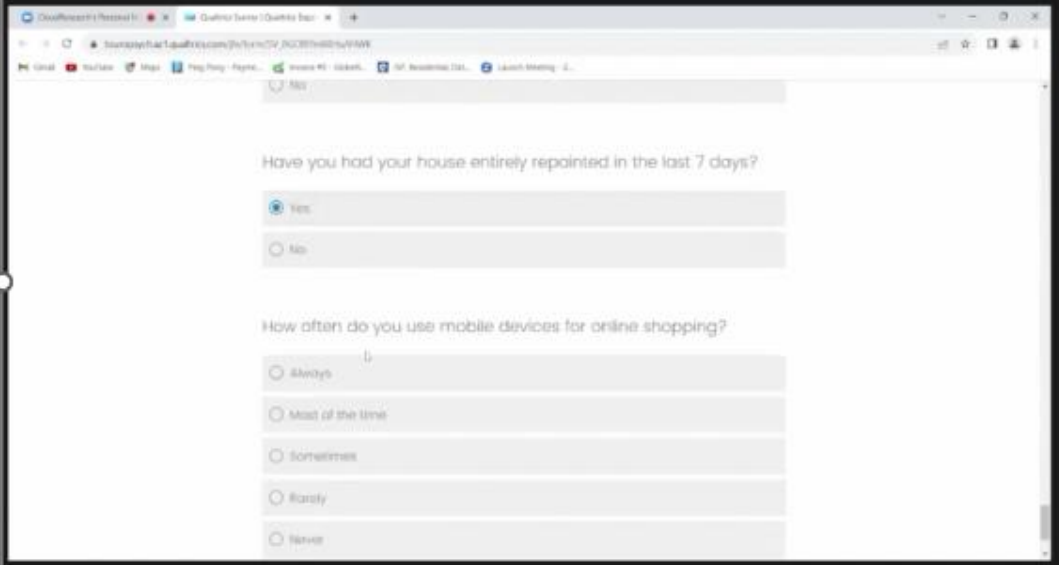
Race	Num.	Perc.
Black or African American (non-Hispanic)	28	87.5%
Other/Mixed	2	6.25%
Hispanic or Latino	2	6.25%

As of the 2020 United States census, there were 32 people, 6 households, and 0 families residing in the town.



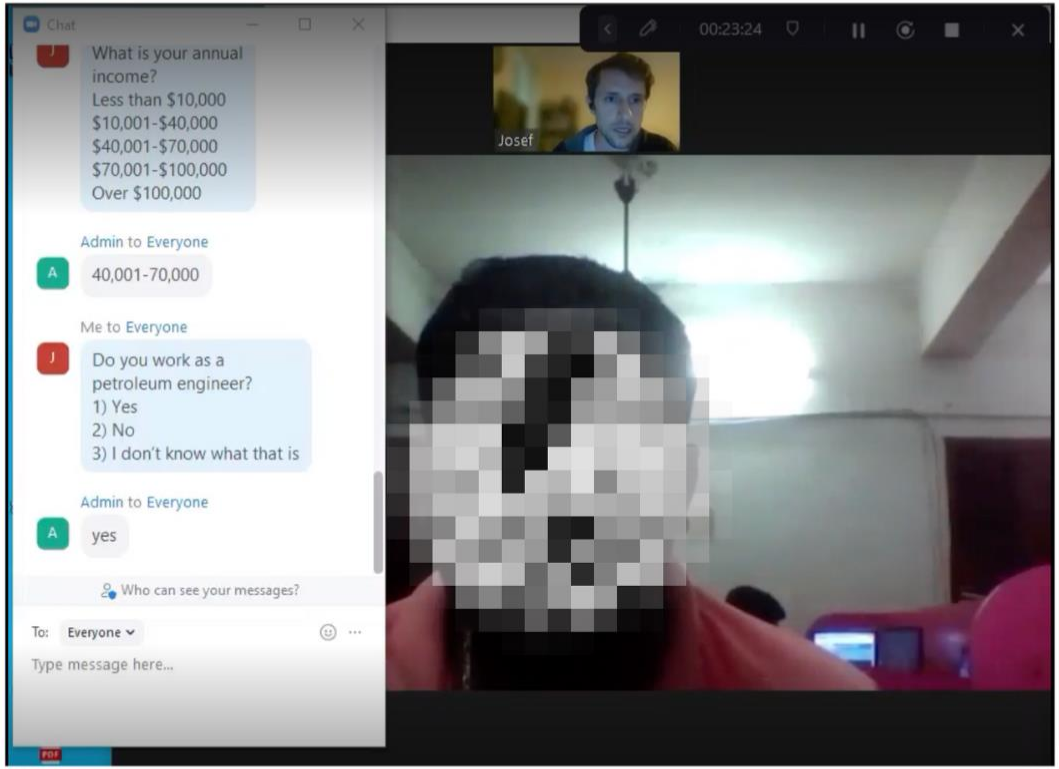
Have you recently purchased a home in McMullen, Alabama?
Population = 32


None of the above
Please select which of the following hair-care products you have used in the last 6 months:
None of the above
Have you visited any webpage on the internet in the last 90 days?
RAM
00:04.64


Have you had your house entirely repainted in the last 7 days?
Yes
No
How often do you use mobile devices for online shopping?
Always
Most of the time
Sometimes
Rarely
Never


RAM

00:01.11


Chat
What is your annual income?
Less than $10,000
$10,001-$40,000
$40,001-$70,000
$70,001-$100,000
Over $100,000
Admin to Everyone
40,001-70,000
Me to Everyone
Do you work as a petroleum engineer?
1) Yes
2) No
3) I don't know what that is
Admin to Everyone
yes
Who can see your messages?
To: Everyone
Type message here...
00:23:24
Josef

Global Network of Fraud

Evidence of such fraudulent activity is easily observed online, including on social media platforms as YouTube, Facebook, Telegram, and Reddit.

“Survey Help 360” based in Bangladesh—44,000 followers on YouTube and 6,000 on Facebook.

- Daily videos explain how to circumvent safety checks on online panels.
- Describe step-by-step how to pose as a U.S. citizen by using rented U.S. cell phone numbers and proxy servers with U.S. IP addresses

“FaijulIslam68” Shows how to create a fraudulent identity to pose as a Black woman using a fake driver’s license and augmented reality so that a potential fraudster can participate in a video survey for Black females .

“Priscy’s Corner” Teaches users how to maximize survey earnings by claiming to fall under multiple target demographics.

- Explains that replying "YES" to as many questions possible increases the likelihood of "qualifying" for a survey.

Source: Litman et al., 2023

Ten Years of Research into the Global Click Farm Industry

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–11

DOI: 10.1177/17456916251404872
www.psychologicalscience.org/PPS

Sage

The Bots Ruining Social Science Are Not Bots at All

Shalom N. Jaffe[1,2], Aaron J. Moss[1], Rachel Hartman[1], Cheskie Rosenzweig[1,3], Richa Gautam[4], Jonathan Robinson[1,5], and Leib Litman[1,6]

[1]CloudResearch, Queens, New York; [2]College of Psychology and Counseling, Fairleigh Dickinson University; [3]Department of Clinical Psychology, Columbia University; [4]Psychological and Brain Sciences, University of Delaware; [5]Department of Computer Science, Lander College; and [6]Department of Psychology, Lander College

Abstract

Researchers who employ online data collection from human subjects currently face a conundrum: It is both essential to how behavioral science functions and threatened by low-quality data. It is often assumed that random, inconsistent, and otherwise incomprehensible data in online surveys comes mainly from bots. Despite this assumption, few studies have directly examined where problematic data comes from, even though identifying the source has important implications for creating the right solutions. We examined this issue on several popular participant-recruitment platforms, including

Source: Jaffe, S. N., Moss, A. J., Hartman, R., Rosenzweig, C., Gautam, R., & Robinson, J. (2026). The bots ruining social science are not bots at all. Perspectives on Psychological Science. Advance online publication. https://doi.org/10.1177/17456916251404872

CloudResearch®

Two Layers. One Answer.

Behavioral detection at every stage of the survey — before it starts, and throughout.

LAYER 1 SENTRY® DATA QUALITY GOLD STANDARD

The Pre-Survey Gate

30–40 seconds.

- Mouse movements, typing cadence, click patterns
- Copy-paste detection in open-ends
- Translation app detection via screen recording
- 300M+ human event library — ML classifier trained on real behavior
- Catches bots, AI agents, and inattentive humans before they enter

LAYER 2 engage®

The Full Survey Platform

Behavioral monitoring doesn't stop at the gate. It runs through the entire survey.

- Continuous event streaming throughout the full survey
- Detects mid-survey LLM use and copy-paste on open ends
- Attention check analysis and response pattern scoring
- Those who make it through the initial layer but then become inattentive or start using LLMs can be caught through more thorough full survey monitoring.

Together: pre-survey + within-survey behavioral detection. The most complete defense available. CloudResearch®

Key Takeaways

Traditional Detection is Insufficient for AI
Standard attention checks, logic puzzles, and red herrings are not effective—AI agents now pass them at a 99.8% rate.

The Economics Are Shifting Rapidly
While human fraud is currently more cost-effective than autonomous agents, this barrier is temporary—AI-driven fraud will soon become scalable and economically viable.

Shift from "What" to "How"
To distinguish humans from AI, move beyond analyzing what participants answer—focus on how they answer: mouse trajectories, keystroke dynamics, and velocity profiles.

Event Streaming Achieves >99% Accuracy
In blind validation testing against sophisticated AI agents, Engage's event streamer system achieved an integrated detection score of >99% with a false positive rate below 1%.

The "Cat and Mouse" Requires Evolution
There is no permanent solution to AI fraud—defense requires a Red Team / Blue Team approach where new attacks constantly test and refine detection algorithms.

Deterrence is Key to Prevention
Combining real-time behavioral flagging with verified participant pools and permanent ban policies creates a risk-reward structure that makes AI-based fraud economically unviable.