

**David Rothschild**  
**Microsoft Research**

# Surveys and AI

**David Rothschild**  
**Microsoft Research**

## **Questions**

***Behavioral Economics & Finance:*** judgement & decision making and market design

***Political Science:*** public opinion trends and predictions

***Marketing:*** Paid (adverts) media

***Communications:*** Earned (news) media

## **Data (Active)**

Generative Models  
Market design/data  
Survey design/data

## **Data (Passive)**

Behavioral data  
Administrative data  
Social media data  
Browser/search data

# relevant recent work

- **Papers:** AI & non-surveys on productivity, work, society/economy
- **Practitioner:** I run surveys for academic & practical use
- **Forthcoming Research:** Summarization of Open Text Responses (3), Representation in Labelling for AI, Data Quality/Catching Bots
- **Papers on AI & Surveys:**
  - “Opportunities and risks of LLMs in survey research” (with James Brand, Hope Schroeder, Jenny Wang)
  - “Successfully Navigating the Disruption AI will Bring to Survey Research” (with Trent Buskirk, Stephanie Eckman, Sunshine Hillygus, Frauke Kreuter, David Lazer)
- **Chair:** “AAPOR Task Force on Responsible AI Integration in Survey Research”

# non-probability data

## Forecasting elections with non-representative polls

Wei Wang<sup>a,\*</sup>, David Rothschild<sup>b</sup>, Sharad Goel<sup>b</sup>, Andrew Gelman<sup>a,c</sup>

<sup>a</sup> Department of Statistics

<sup>b</sup> Microsoft Research

<sup>c</sup> Department of Political Science

Columbia University, New York, NY, USA

New York, NY, USA

*Quarterly Journal of Political Science*, 2016, 11: 103–130

## The Mythical Swing Voter

Andrew Gelman<sup>1</sup>, Sharad Goel<sup>2</sup>, Douglas Rivers<sup>3</sup> and David Rothschild<sup>4\*</sup>

<sup>1</sup> Columbia University, USA; [gelman@stat.columbia.edu](mailto:gelman@stat.columbia.edu)  
<sup>2</sup> Stanford University, USA; [sgoel@stanford.edu](mailto:sgoel@stanford.edu)  
<sup>3</sup> Stanford University, USA; [rivers@stanford.edu](mailto:rivers@stanford.edu)  
<sup>4</sup> Microsoft Research, USA; [davidmr@microsoft.com](mailto:davidmr@microsoft.com)

## Non-Representative Surveys: Fast, Cheap, and Mostly Accurate\*

Sharad Goel  
Stanford University

Adam Obeng  
Columbia University

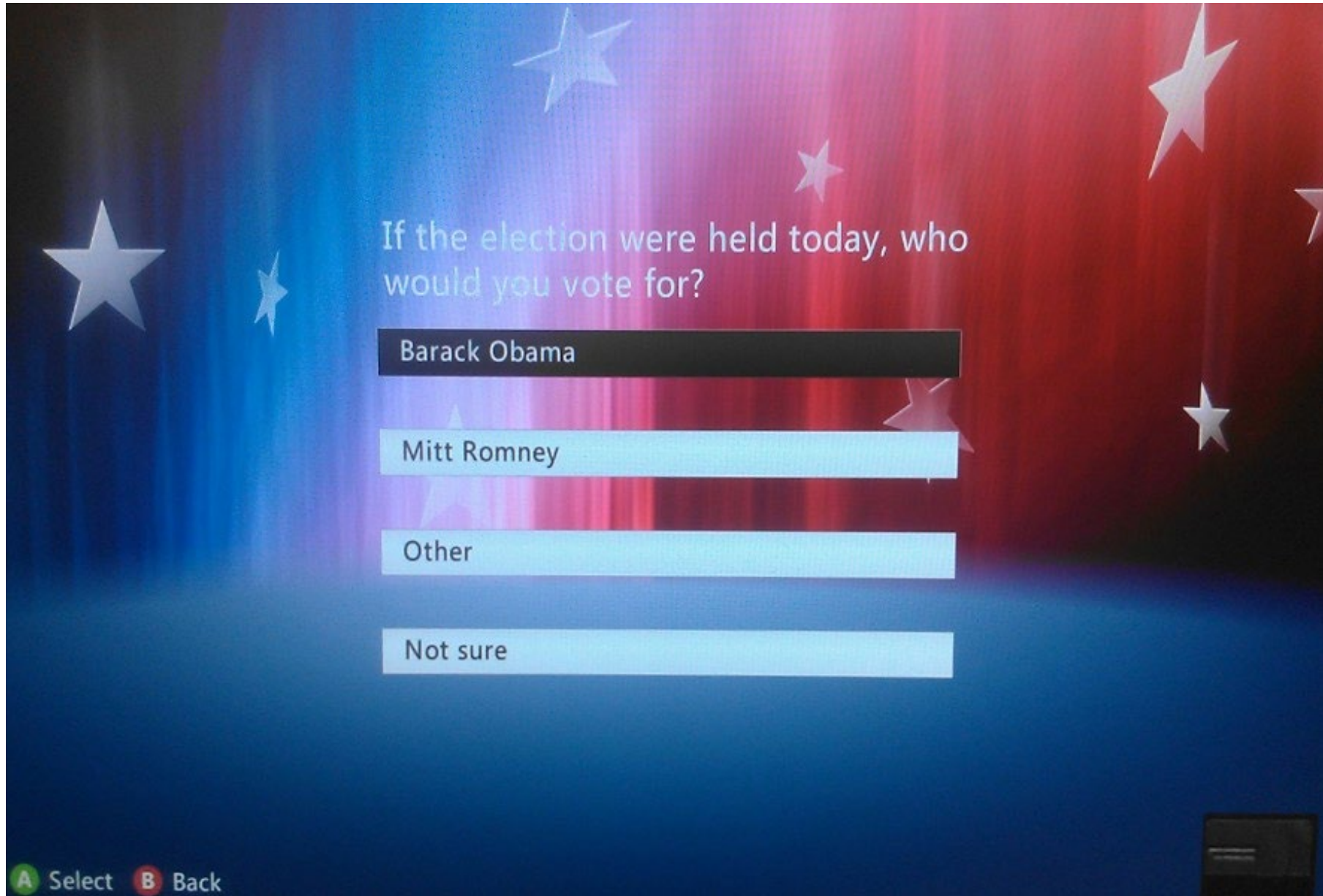
David Rothschild  
Microsoft Research

## Online and Social Media Data As an Imperfect Continuous Panel Survey

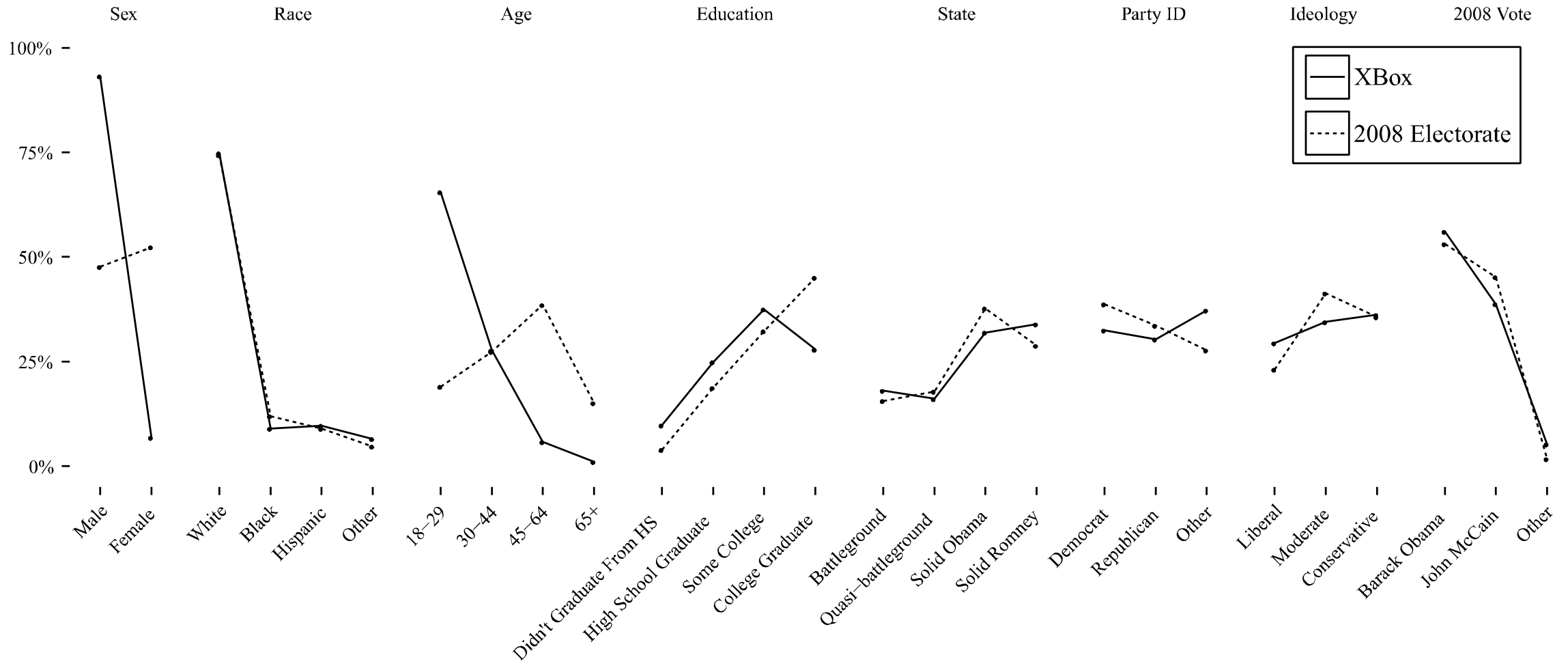
Fernando Diaz , Michael Gamon , Jake M. Hofman , Emre Kiciman , David Rothschild  

Published: January 5, 2016 • <https://doi.org/10.1371/journal.pone.0145406>

# non-probability data

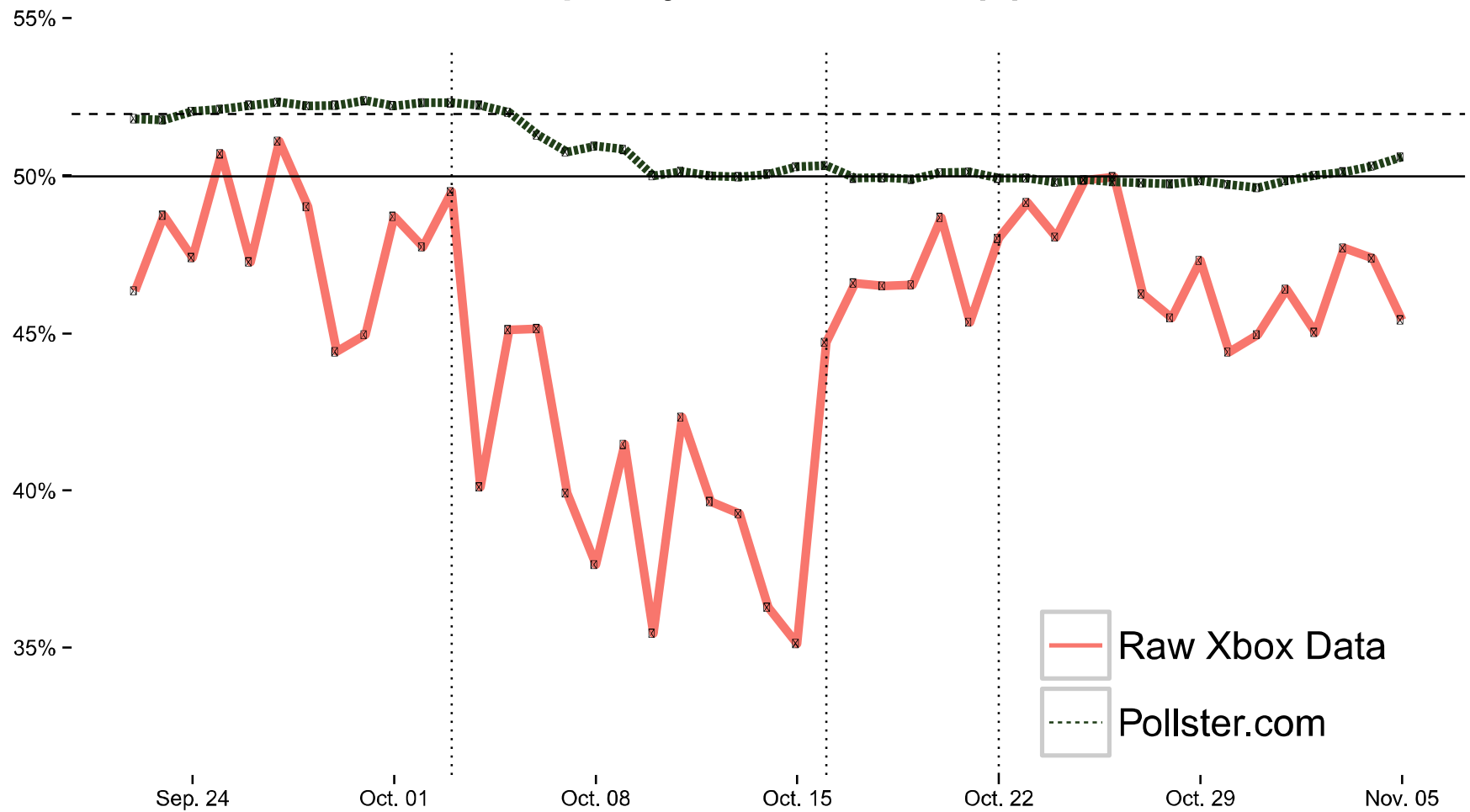


# non-probability data

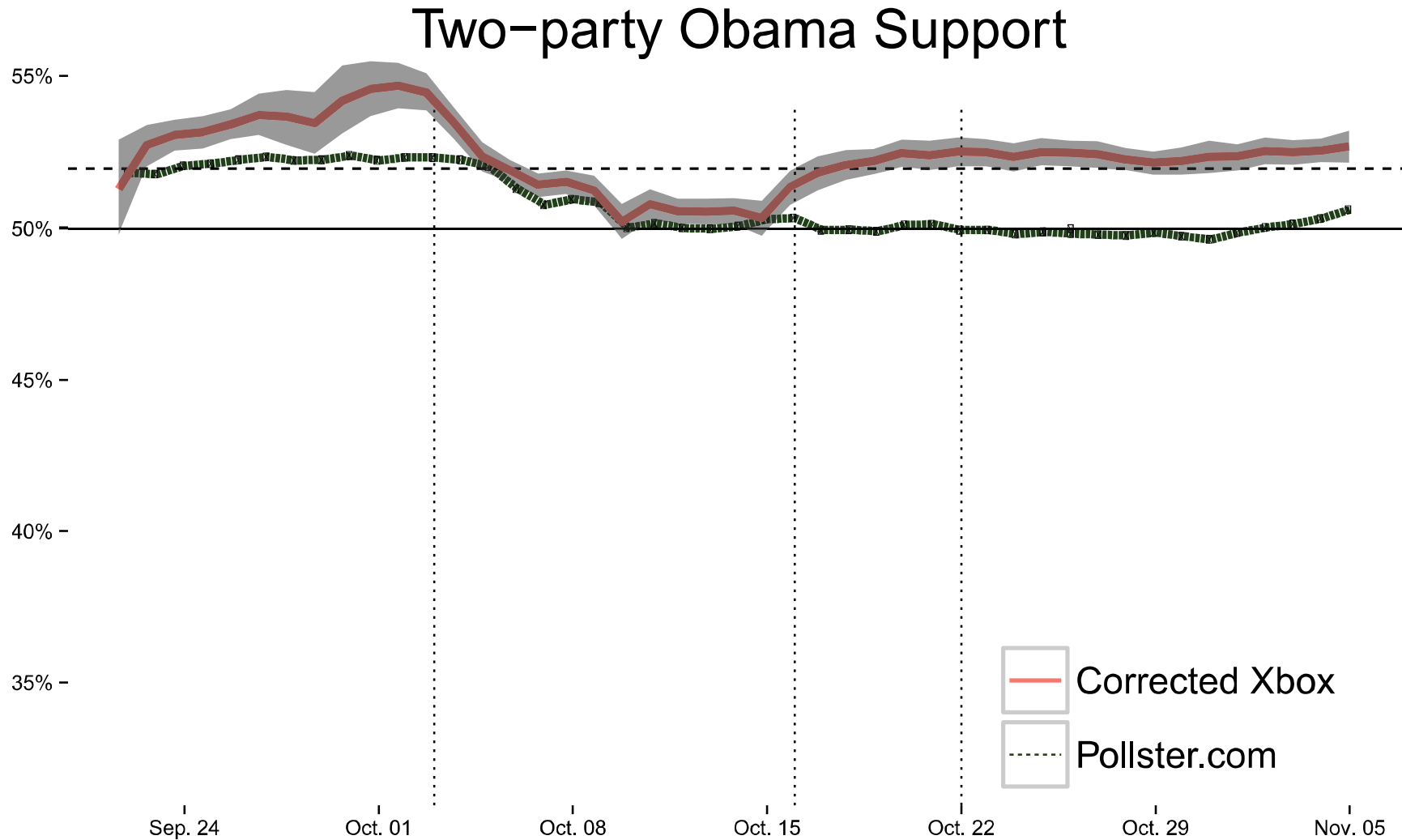


# Xbox Panel: Raw Response

## Two-party Obama Support



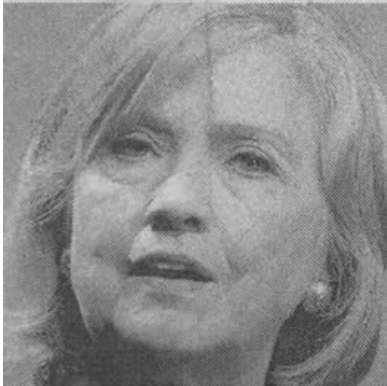
# Xbox Panel: Response with MPR



## **TheUpshot**

### **How Four Pollsters, and The Upshot, Interpreted 867 Poll Responses**

We gave four good pollsters the same raw data behind a recent New York Times Upshot/Siena College poll in Florida. They had four different results. Below, their estimates, along with The Upshot's.



**Clinton +3**

Charles Franklin  
Marquette Law



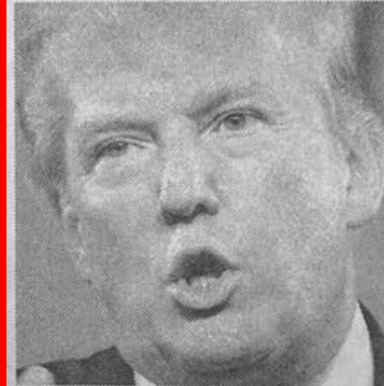
**Clinton +1**

Patrick Ruffini  
Echelon Insights



**Clinton +4**

Margie Omero,  
Robert Green,  
Adam Rosenblatt  
  
Penn Schoen  
Berland Research



**Trump +1**

Sam Corbett-Davies,  
Andrew Gelman,  
David Rothschild  
  
Stanford University,  
Columbia University  
and Microsoft Research

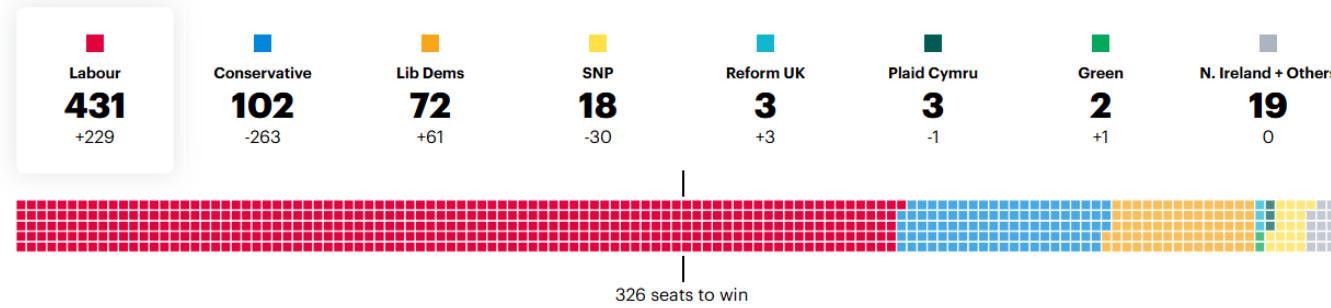


**Clinton +1**

New York Times  
Upshot/  
Siena College

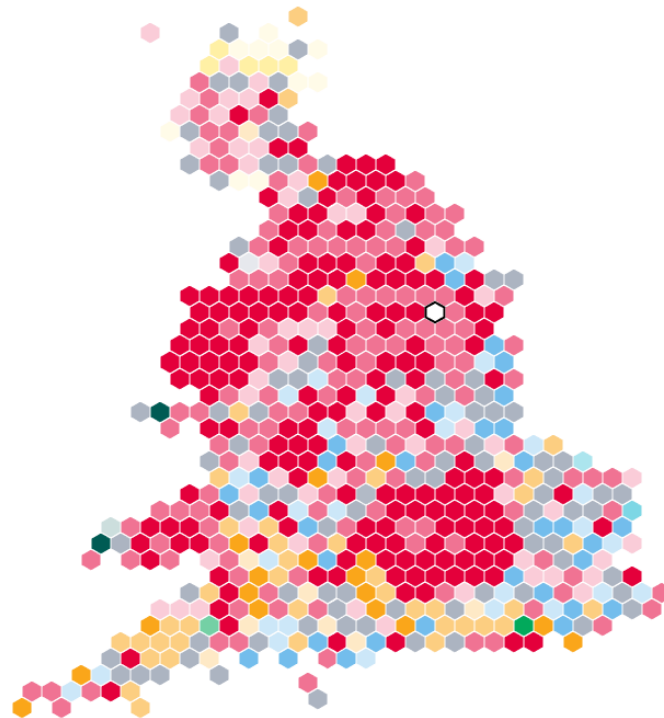
# Our latest MRP seat projection ⓘ

Last update: June 19, 2024 — July 2, 2024



Show seat changes

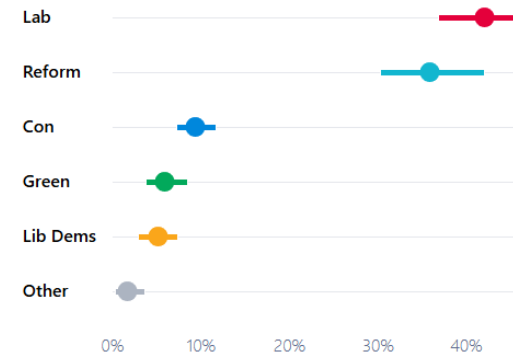
🔍 Search postcode, constituency or candidate



## Barnsley North

Lean Labour

Hold Labour

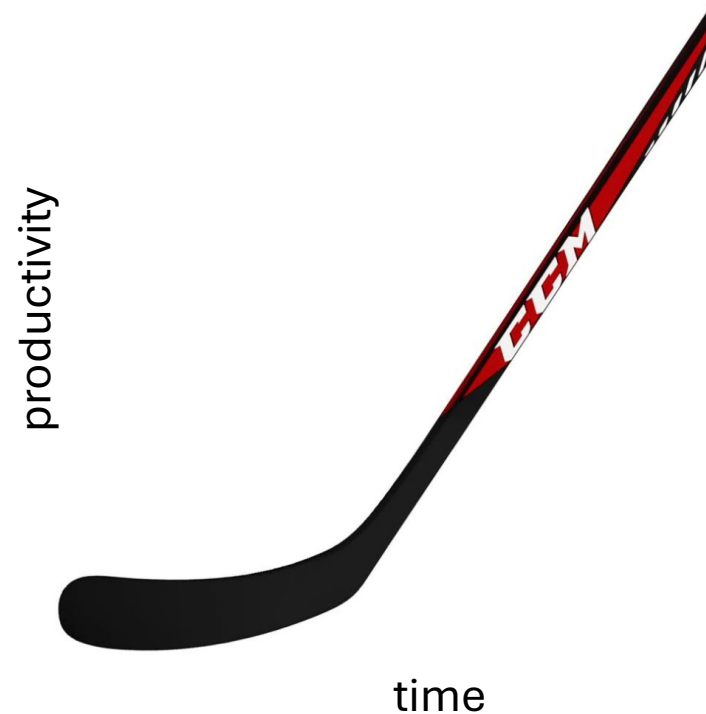


# non-probability disruption

- Expanded access to surveys
- Left industry fragmented: costly for industry built on trust
- Set baseline for next disruption

# stages of disruption

- (1) Augmentation** of tasks within existing workflows
- (2) Redistribution** of resources within existing workflows
- (3) Reconstruction** of workflows



# augmentation (early results)

- Generally, people choosing speed over quality upgrade (enjoying the diminished effort)
- Benefits higher for low skilled in Global North, but high skilled in Global South
- Mixed results on augmentation v. replacement & deskilling



Sign In

Subscribe

AI And Machine Learning

## A Sports Analogy for Understanding Different Ways to Use AI

by Jake M. Hofman, Daniel G. Goldstein and David M. Rothschild

December 4, 2023



# disruption

- ***Disruption is reimagining workflows to fully leverage AI's capabilities:*** not augmenting current workflows
- ***AI's advance is inevitable, but its societal trajectory is not:*** imperative for thoughtful individuals and institutions to actively shape its trajectory toward openness, innovation, and broad-based welfare gains
- Surveys can become so much more powerful, more central: scalable, qualitative-enhanced, linked-panels, but need:
- Survey researchers to force transparency, standards, collaboration on models and tools

# disruption

- ***Disruption is reimagining workflows to fully leverage AI's capabilities:*** not augmenting current workflows
- ***AI's advance is inevitable, but its societal trajectory is not:*** imperative for thoughtful individuals and institutions to actively shape its trajectory toward openness, innovation, and broad-based welfare gains
- Surveys can become so much more powerful, more central: scalable, qualitative-enhanced, linked-panels, but need:
- Survey researchers to force transparency, standards, collaboration on models and tools

**Augmentation not Replacement:** Surveys flourish and Survey researcher flourish

# AI (LLMs) dimensions

- Model
- Access/Tooling
- Prompts/Instructions

# survey pipeline (stylized)

- (0) **Measure Ideation:** create preliminary questions for measures
- (1) **Survey Ideation:** Researchers turn preliminary questions into tractable survey questions
- (2) **Administration:** Researchers launch a survey instrument to the target audience.
- (3) **Data Analytics:** Researchers clean the data and analyze the results.
- (4) **Reporting to Stakeholders:** Researchers convey the results back to relevant stakeholders.

# survey pipeline (stylized total survey error)

- (0) **Measure Ideation:** specification error
- (1) **Survey Ideation:** measurement error
- (2) **Administration:** sampling & coverage & non-response error
- (3) **Data Analytics:** processing error
- (4) **Reporting to Stakeholders:** \*reporting error\*

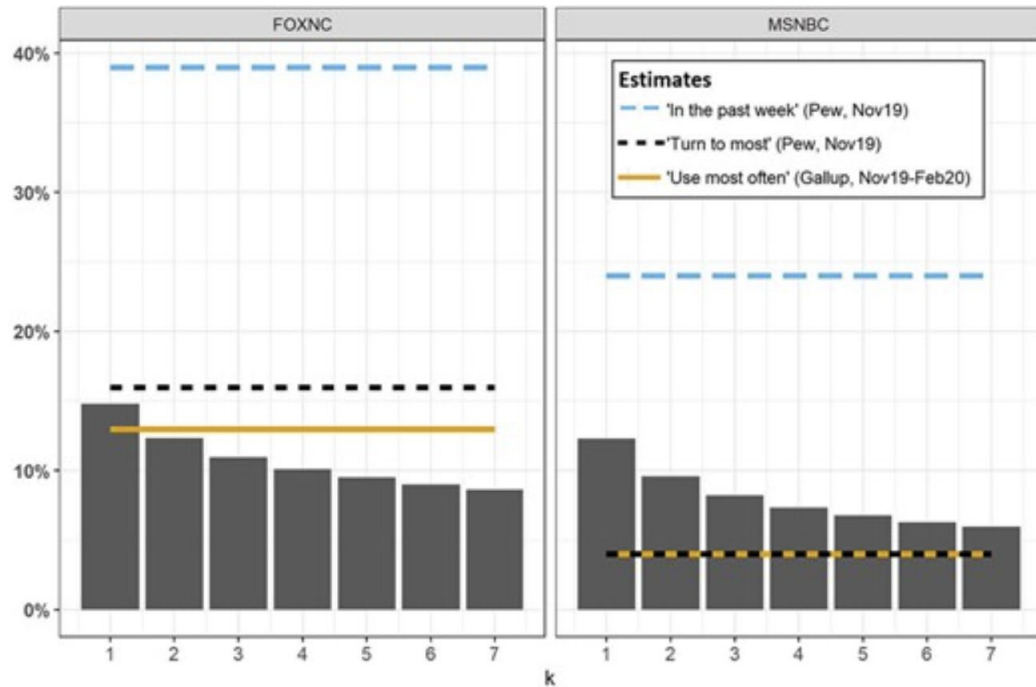
# categorization

- Stages of disruption
- Placement in current workflow
- Agency for principal
- Risk v. reward
- Dimensions of interest
- Qualitative v. quantitative
- Global North v. South
- Blockers: methods, tooling, norms, incentives

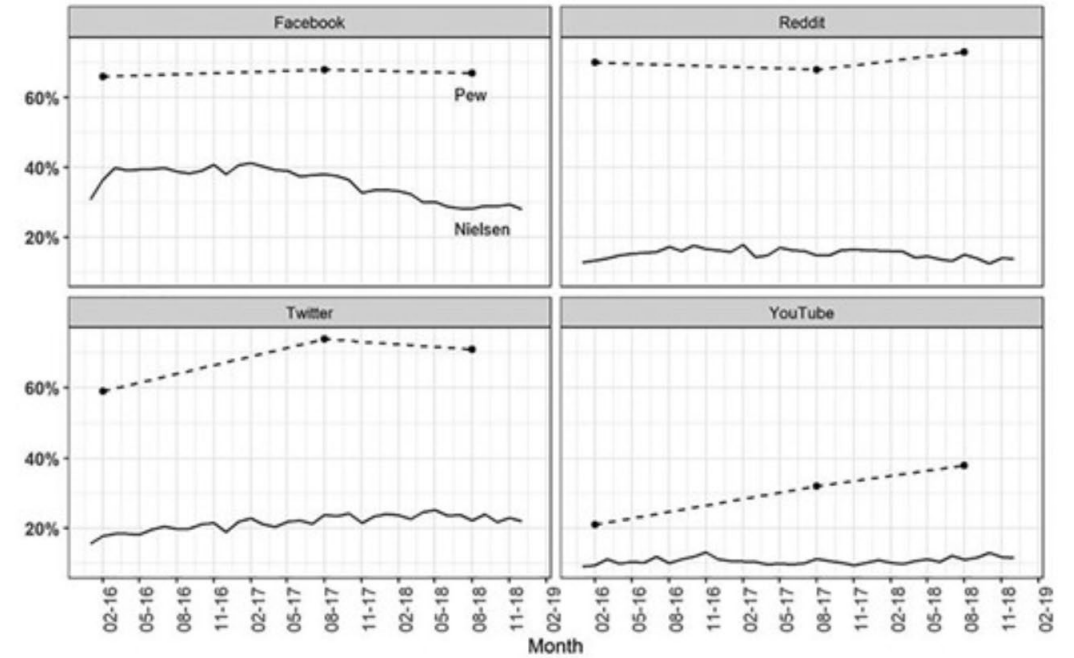
# (0-1) pre-data: measure and survey ideation

- **Editing** surveys/questions (lower risk/reward, higher agency)
- **Generating** surveys/questions (higher risk/reward, lower agency)

# (0-1) pre-data: editing questions



**Figure 1.** Comparison of estimates of FOXNC and MSNBC consumption between passively collected and survey sources: percent of US adults watching  $k$  or more sessions (six-minute blocks) of FOXNC or MSNBC. Bars show estimates from passively collected data from Nielsen Television Panel, November 2019.



**Figure 5.** Comparison of news consumption estimates conditional on being on a social media platform. Nielsen: Percentage of platform users consuming one (or more) news articles linked from social media platforms, by platform and month. Pew data from News Use Across Social Media Platforms 2016, 2017, 2018.

# (0-1) pre-data: editing questions

How often in the last week did you hear about politics, or discuss politics, with friends and family?

How often did you talk politics with your friends and family in the last week?

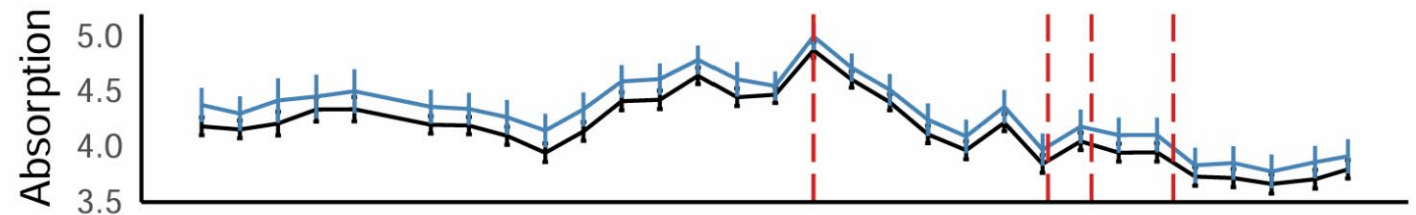
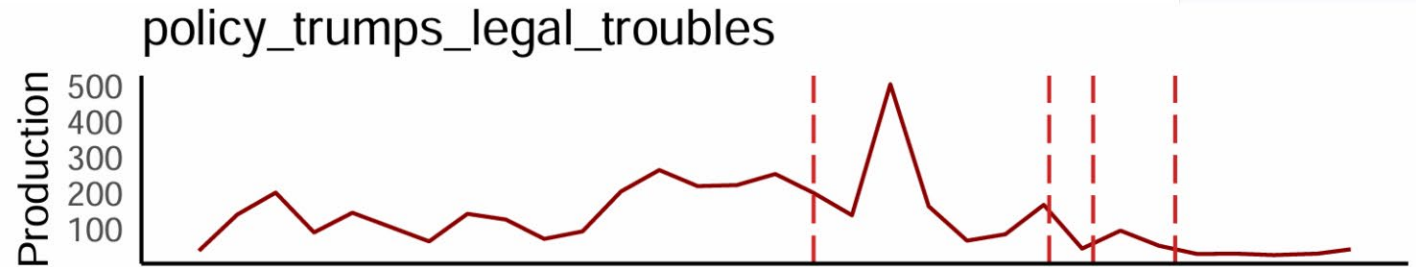
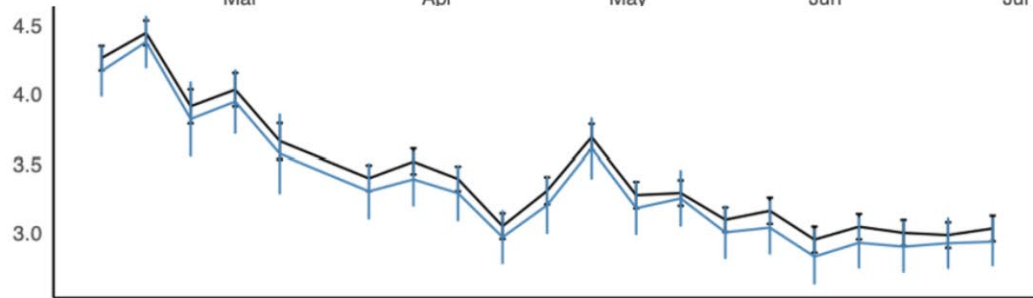
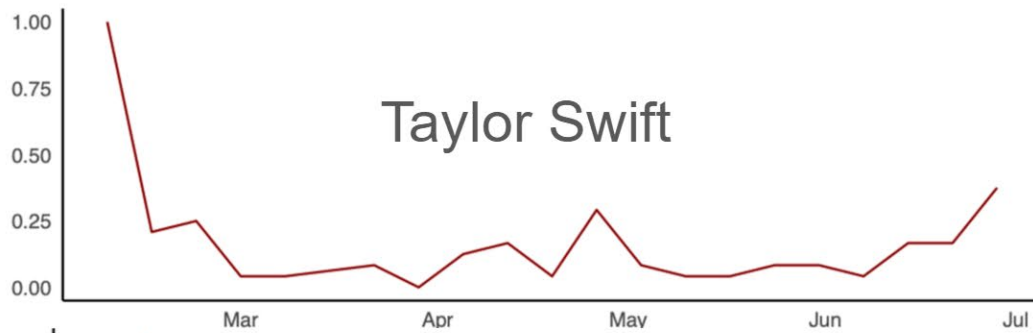
SELECT ONE

- Multiple times a day
- Daily
- Once or twice
- Never



SELECT ONE

- Not at all
- Irregularly / Occasionally
- A few times (1-2 times)
- Several times a week (3-4 times)
- Once a day
- Multiple times a day (2 or more times)



# (0-1) pre-data: generating questions

“There are people in my party who go to Washington to bark, to make noise — not to make law, but to make noise. I think Jim Jordan would call himself one of those, who’s got a lot to say and is loud and barking, but actually passing law, getting law that’s signed, not just by members of the House, but also in the Senate and by the president,” Romney added in the interview. “That’s a different matter, and we’ll see whether he rises to the occasion if he becomes speaker.”

While he [ranks low](#) in terms of the number of bills he has introduced over his nine-term tenure — [none of which](#) have become law — Jordan has made a name for himself among conservative pundits and grassroots activists by playing a key role in the government shutdowns of 2013 over Obamacare and 2018 over a border wall and spearheading the investigation into President Joe Biden’s business dealings.

After Jordan’s failure to win over a majority of his conference Tuesday, he continued to try to win over his conference members, with another possible vote coming on Wednesday.

“He’s probably not the first choice I would’ve made, but it’s not my choice, it’s up to them,” Romney said. “But if he does get the job, it’ll be a case of the dog catching the car, which is, what happens then?”

## NEWS POLL

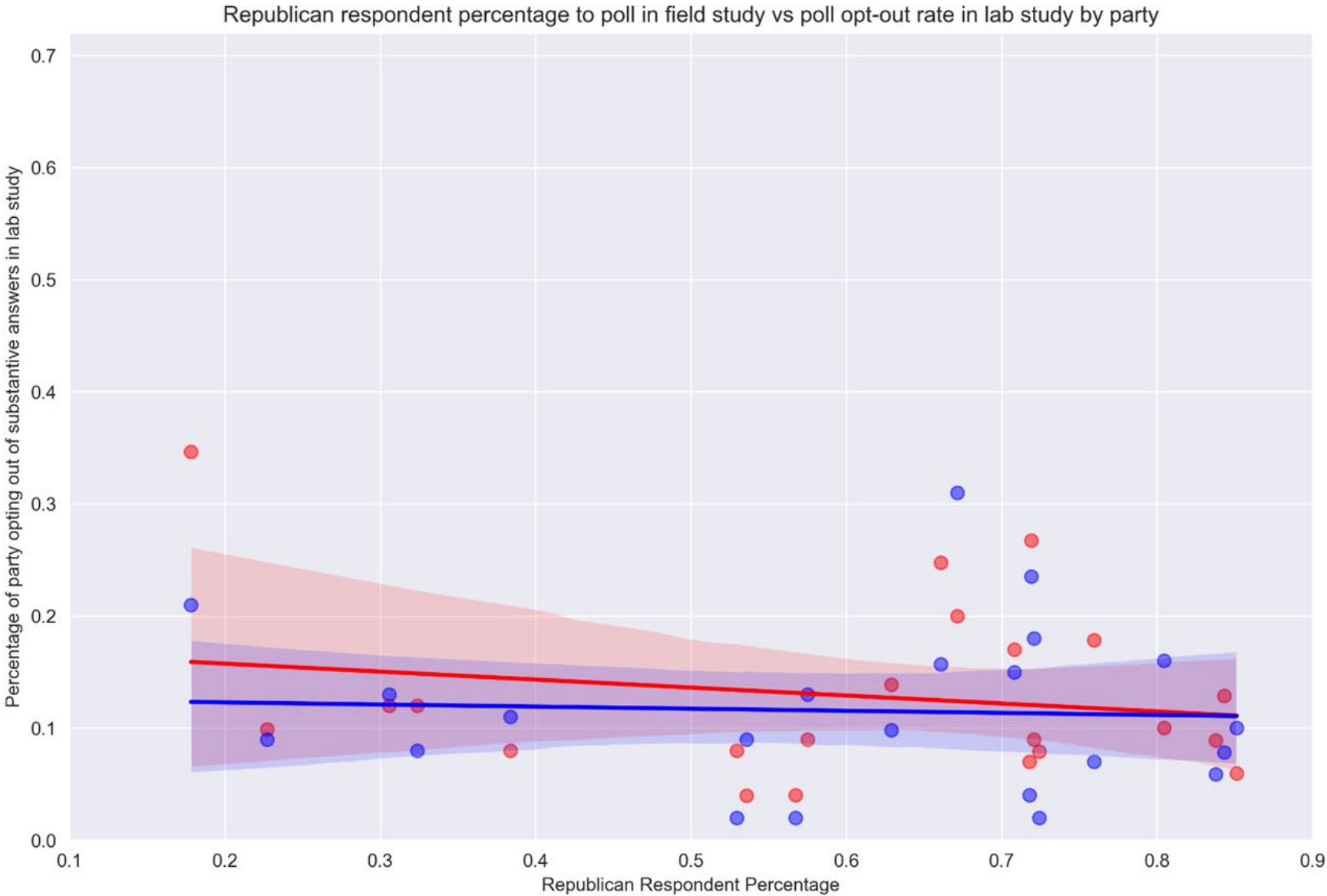


**Which of the author's points in this article about the House Speaker race are you likely to remember in a month?**

- Utah's Republican senators Lee and Romney expressed differing views on whether House Republicans should unite behind Rep. Jim Jordan for speaker.
- Jordan didn't secure a vote majority to become speaker Tuesday, with some expressing concerns over his role in rejecting 2020 election results.
- Lee expressed support for Jordan, while Romney said he wasn't the right choice for House speaker.
- Jordan has made a name for himself among conservative pundits by spearheading investigations into President Joe Biden's business dealings.
- I don't have a takeaway from this article.
- My takeaway from this article is not in this list.

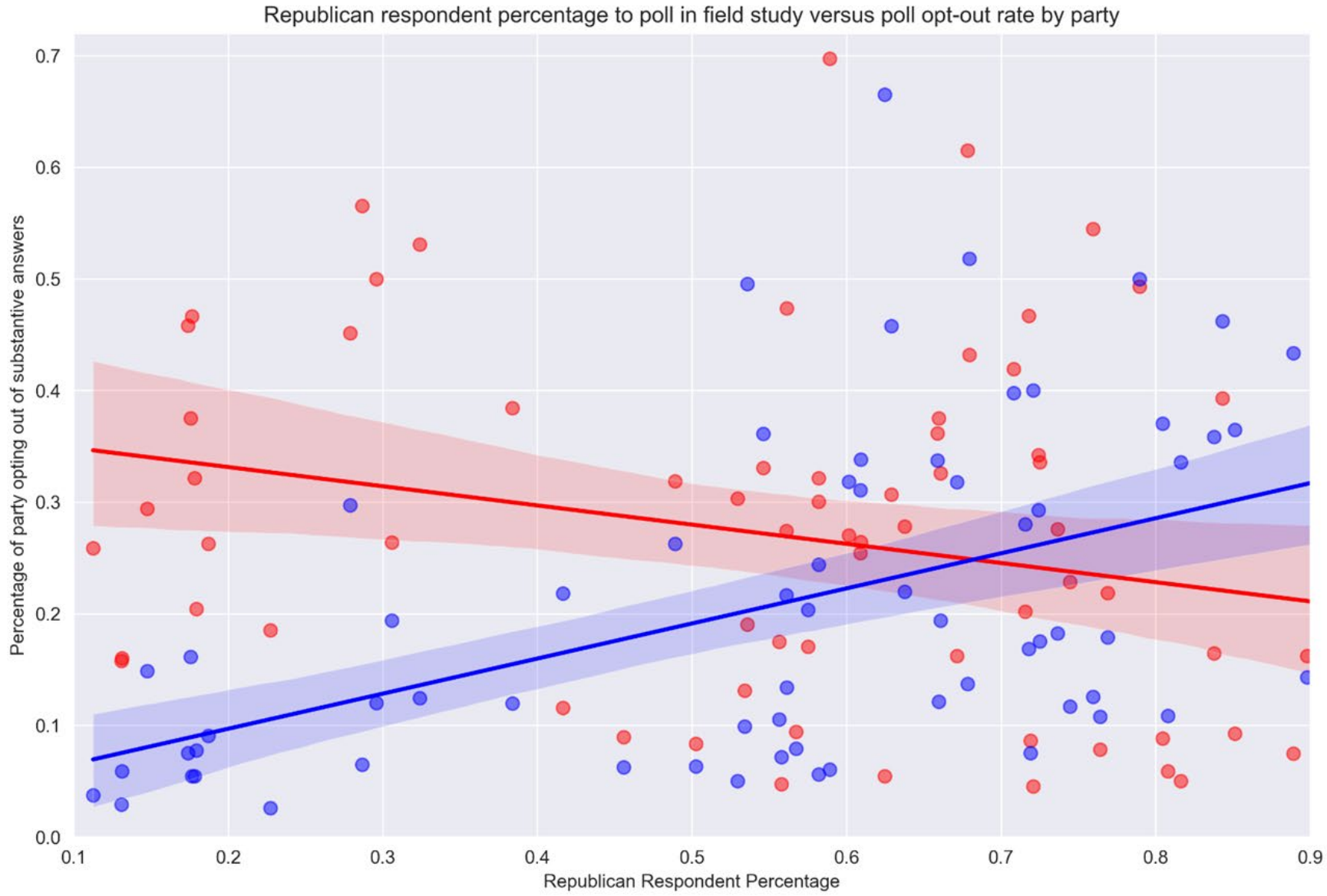
NEXT

# Lab: non-substantive answers are not correlated with lean of respondents



Field: non-substantive answers are correlated with lean of respondents

Asymmetric: Republicans opt-out at much higher rates



# (0-1) pre-data: research/risks/opportunity

- **Training Data:** What data is a given model drawing on for their answers? How do the outputs improve (or get worse) with fine-tuning?
- **Across Domain:** How do LLMs perform across domains (e.g., political attitudes vs. consumer preferences)?
- **Behavioral Bias:** Do different uses of LLM-based tools, such as question generation versus critique, introduce distinct behavioral biases or foster over-reliance?
- **Prompts:** How robust are the outputs to prompts and how can methods or tooling improve the output?
- **Experts v. Non-Experts:** Can tools adjust for expertise (where measurement errors are likely quite different)?
- **All-up Ideation:** Can tools induce the right feedback loop between substantive expertise and survey methodology (plausibly shifting how we understand and address specification error)?

## (2a) AI as the interviewer

- **Risks (high):** consistency, bias, transparency, danger
- **Rewards (high):** increase engagement, reduce respondent fatigue, improve data richness

# Smarter Surveys. Deeper Insights.

Voiceform combines structured surveys with voice, video, and AI to uncover deeper insights in minutes.



**CROWD VOICE**

The power to **hear the voice of your people**, with massive interview campaigns via WhatsApp and AI. The intelligence to understand their **opinions, behaviors & needs**, and to **crowd-source** their ideas and solutions.

# (2a) AI as the interviewer: research/risks/opportunity

- **Safeguards:** What safeguards are needed to protect against bias or harm?
- **Documentation:** What forms of documentation or logging are required to make AI-mediated interactions transparent and replicable?
- **Respondent Perception:** How do respondents perceive AI interviewers compared to human interviewers with respect to trust, response accuracy, and willingness to participate?
- **Mode:** How does the choice of mode (text or voice) affect responses?
- **Fatigue:** Does the increased use of open-ended questions increase respondent burden and reduce overall data quality and usefulness?
- **Attack:** How does AI defend against attacks or long-tail spirals?

## (2b) AI as the respondent

- **Off-the-shelf Issues:** fine-tuning model improve accuracy
  - **Robustness:** many ways to do this, many sets of answers. Just ask question, ask question by demographics, repeat question with higher temperature, pull distribution of token probabilities, etc.
  - **Complexity:** LLMs do not like complexity!
  - **Accuracy Varies:** “Errors” higher as you move away in time, population, question
  - **Cost Savings:** Internet and non-probability already dramatically lowered cost of human respondents
- ⇒ High risk, Lowish reward (for now)



## ▲ Launch HN: Roundtable (YC S23) – Using AI to Simulate Surveys

121 points by timshell on July 25, 2023 | [hide](#) | [past](#) | [favorite](#) | [91 comments](#)

Hi HN, we're Mayank and Matt of Roundtable (<https://roundtable.ai/>). We use LLMs to produce cheap, yet surprisingly useful, simulations of surveys. Specifically, we train LLMs on standard, curated survey datasets. This approach allows us to essentially build general-purpose models of human behavior and opinion. We combine this with a nice UI that lets users easily visualize and interpret the results.

Surveys are incredibly important for user and market research, but are expensive and take months to design, run, and analyze. By simulating responses, our users can get results in seconds and make decisions faster. See <https://roundtable.ai/showcase> for a bunch of examples, and <https://www.loom.com/share/eb6fb27acebe48839dd561cf1546f131> for a demo video.

# Behavior doesn't lie.

Roundtable is building the identity layer for digital trust. Our core thesis is that identity is behavioral—a continuous, dynamic fingerprint shaped by action, not static credentials.

Fraudsters can fake credentials, spoof devices, and steal passwords. But they can't fake behavior at scale. We build models that separate real from fake by analyzing session dynamics, cognitive patterns, and interaction flows—verifying identity in real-time, without friction.

Book a demo



## (2b) AI as the respondent: research/risks/opportunity

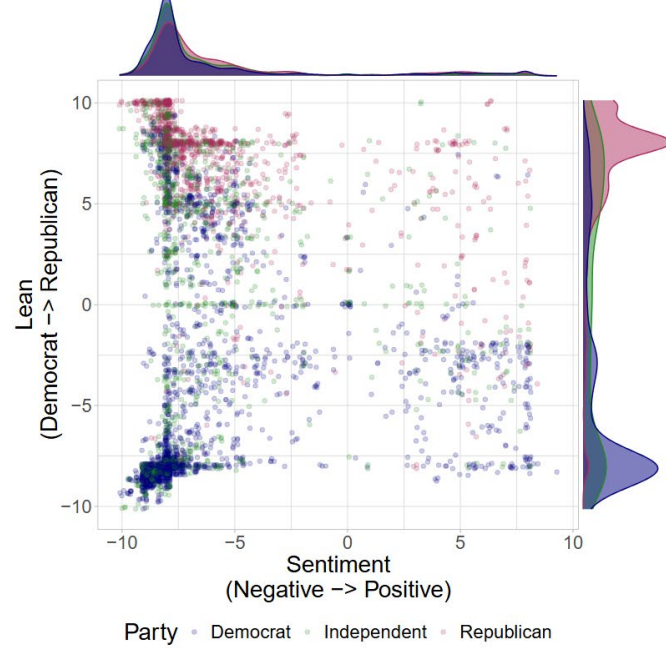
- **Benchmark Across Domains and Populations:** Where synthetic respondents succeed or fail—especially for small sub-populations, marginalized groups, and culturally sensitive topics?
- **Modeling Within-Group Heterogeneity and Correlations:** Can synthetic responses capture intra-group variability and between question correlation?
- **Domain-Specific Fine-Tuning:** Explore the creation of specialized models?
- **Bias Detection and Mitigation in Training Data:** Identify and correct historical, systemic, and behavioral biases in LLM training data to prevent amplification of misrepresentation?
- **Impact of Synthetic Data Contamination:** Study the prevalence and consequences of human respondents using LLMs to complete surveys and develop best practices for detection and prevention.

# (3) AI as a Data Translator/Cleaner/Labeler/Modeler

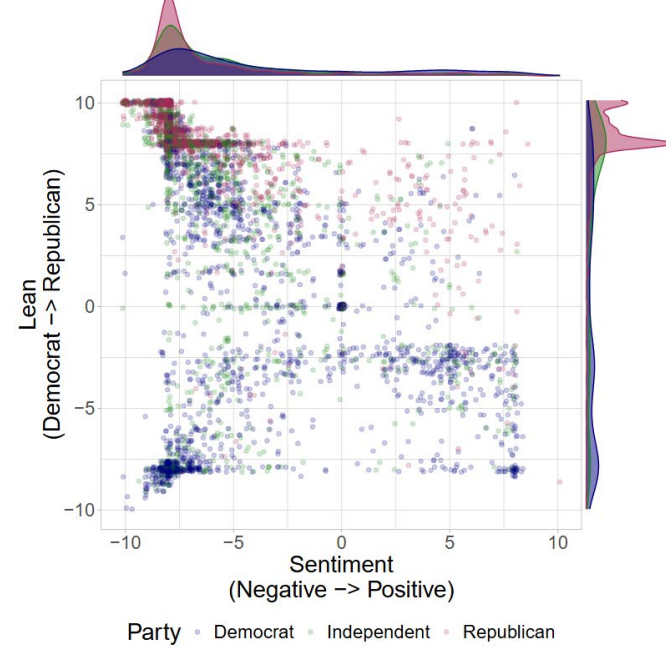
- **Translation/Transcription:** low risk/high reward
- **Data Cleaning:** low risk/low reward
- **Labeling/Summarization:** medium risk/high reward
- **Modelling:** medium risk/high reward

# (3) AI as a Data Translator/Cleaner/Labeler/Modeler

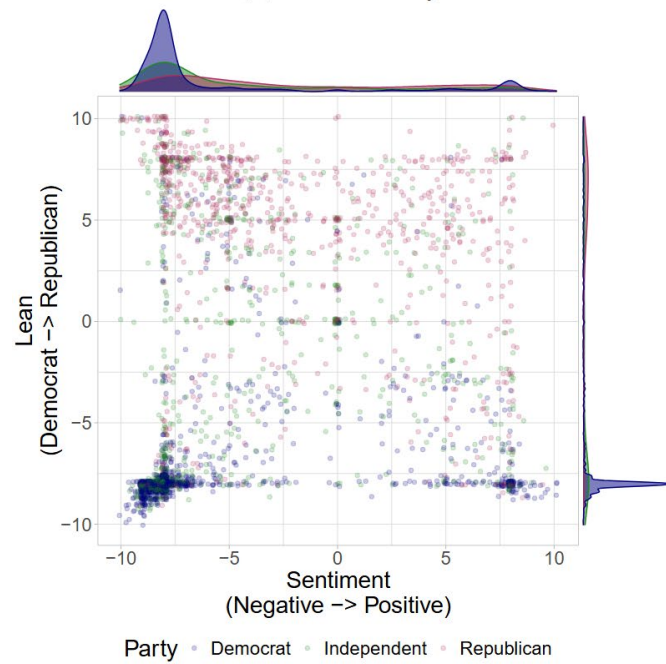
- Prolific on Qualtrics, 2,000 respondents each in 4 waves (September & October 2024), each answering 1 of 4 topic specific narrative questions (so we have 2,000 responses per question below):
  - In 1 or 2 sentences: What is your overarching takeaway that best explains recent events regarding democracy in the US?
  - In 1 or 2 sentences: What is your overarching takeaway that best explains recent events regarding the US economy?
  - In 1 or 2 sentences: What is your overarching takeaway that best explains recent events regarding immigration to the US?
  - In 1 or 2 sentences: What is your overarching takeaway that best explains recent events regarding access to women's healthcare in the US?



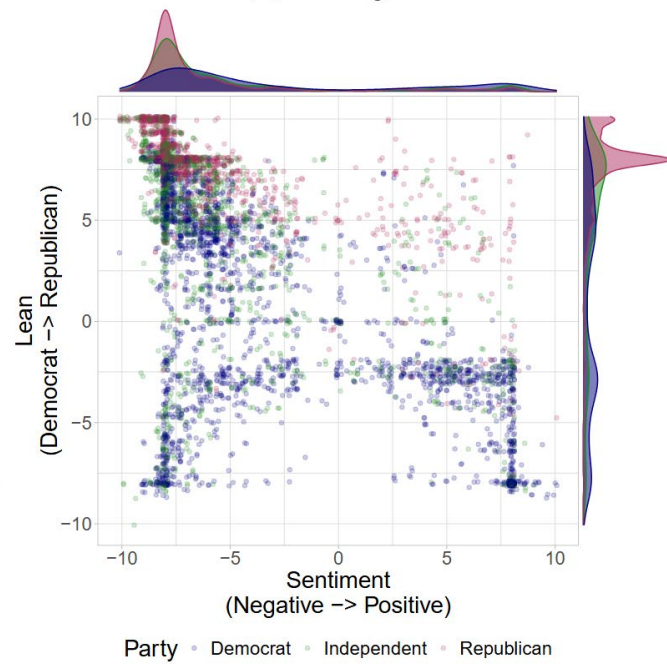
(a) Democracy



(b) Immigration

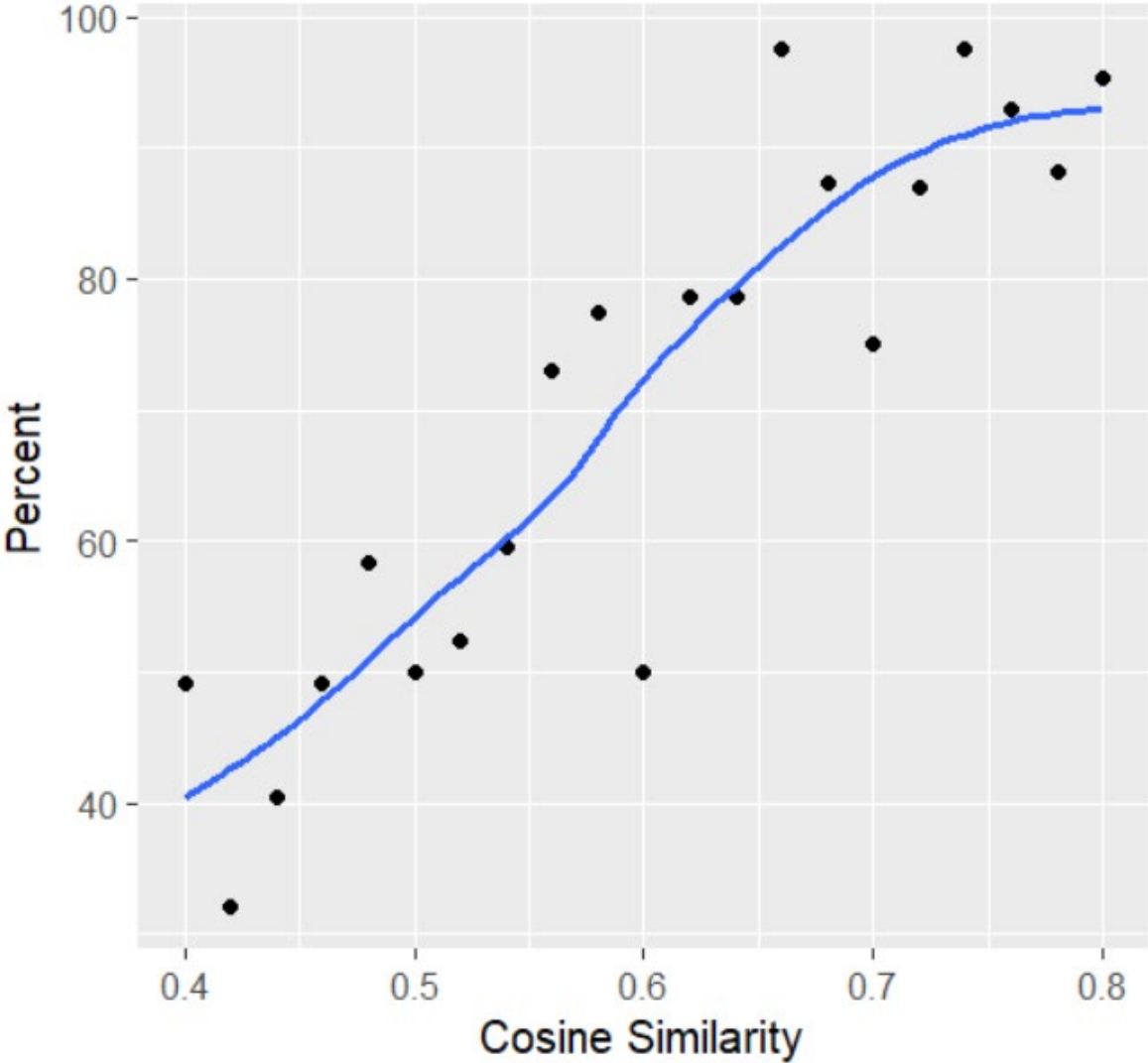


(c) Reproductive Rights



(d) Economy

Are these 10 topics  
derived from open-ended questions  
basically the same?



# (3) AI as a Data Cleaner/Labeler/Modeler - research

- **Stability and Reliability of LLM-Based Labeling:** How do variations in prompts, model versions, and fine-tuning affect the consistency of coded outputs across different topics and populations?
- **Quality Assessment Without Ground Truth:** Explore frameworks for validating AI-generated labels when definitive ground truth is unavailable?
- **Ethical and Privacy Safeguards for Open-Ended Data:** Examine protocols for securely processing open-text responses through commercial LLMs

# AI as Pilot Tester

## **Simulation and Stress-Testing with Synthetic Data:**

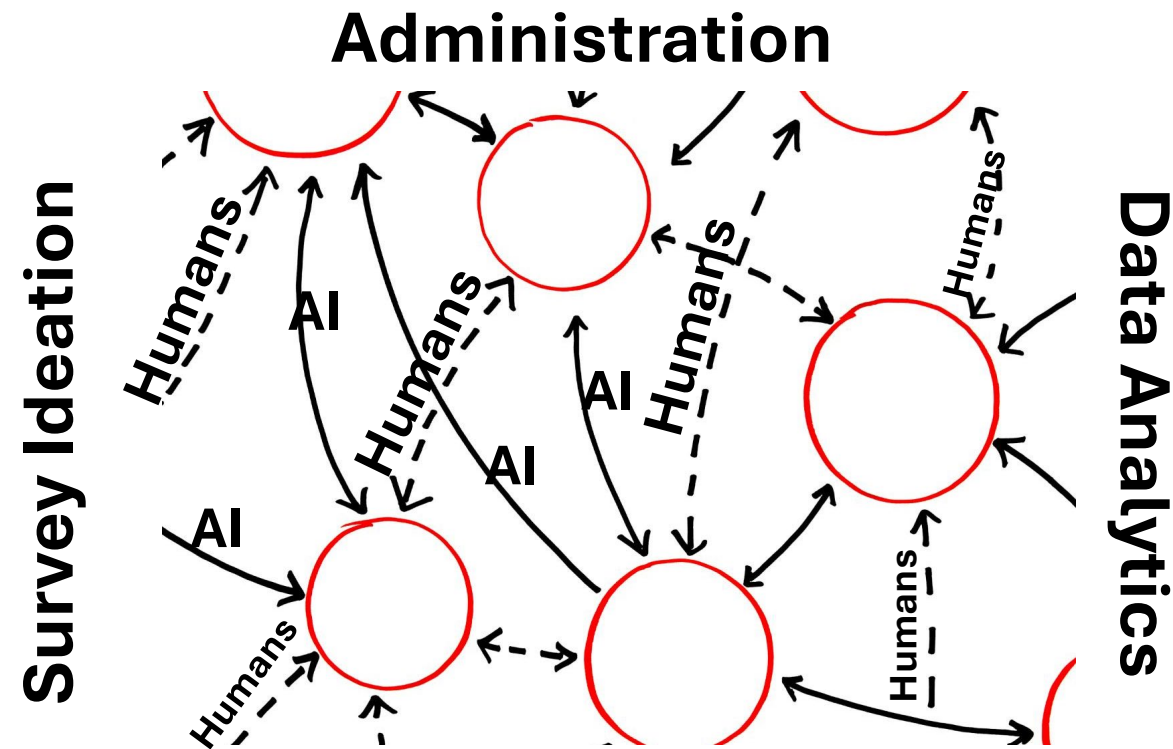
Explore how LLM-generated pilot datasets can be used to validate ***survey questions & logic, coding*** scripts, ***target populations*** and ***statistical models*** before deployment, prepare detailed ***pre-registration***, and assess their fidelity to real-world data structures.

# AI as Interviewer & Respondent & Modeler?

- (1) **Measure Ideation:** create preliminary questions for measures
- (2) **Survey Ideation:** Researchers turn preliminary questions into tractable survey questions
- (3) **Administration:** Researchers launch a survey instrument to the target audience.
- (4) **Data Analytics:** Researchers clean the data and analyze the results.
- (5) **Reporting to Stakeholders:** Researchers convey the results back to relevant stakeholders.

# AI as Interviewer & Respondent & Modeler?

**(1) Measure Ideation:** create preliminary questions for measures




**(5) Reporting to Stakeholders:** Researchers convey the results back to relevant stakeholders.

## (4) AI as a Briefer

- Writing Reports: Medium risk/Medium reward
- RAG (Retrieval-Augmented Generation): Medium risk/High reward


# (5) AI as a Briefer




**Sage**  
Consumer insights assistant  
Full access

- Chat
- Get InsightStore
- Help
- Contact Us
- Manage account
- Logout

[Logo](#)

 **Me** 11:14 AM  
do people trust survey research?

 **Sage** 11:14 AM  
I'll have that for you in a moment.

 **Sage** 11:14 AM

## Summary

In the United States, 4% of consumers completely trust traditional nationwide polls for presidential races, while 30% trust them to a certain extent. However, 61% do not trust these polls at all. An additional 5% have other views or no opinion on the matter.

 **Sage** 11:14 AM

## Details and Insights

In the United States, consumer trust in various institutions and scenarios shows a diverse range of attitudes. Approximately 55% of consumers are not at all likely to trust a company with their personal information following a data breach, while 36% are somewhat likely, and 10% are very likely to maintain trust. Trust in grocery stores is relatively high in states like Delaware (80%) and Wyoming (83%), and older individuals tend to trust poll workers and banks more, with the highest trust levels observed in those 65 or older. Conversely, trust in grocery stores for food safety and traditional nationwide polls for presidential races decreases with age. Only 3% of respondents completely trust traditional nationwide polls for presidential races, while 34% trust them to a certain extent, and a majority of 59% do not trust them at all. Trust in major companies has seen a modest increase over time, and traditional banks are the most trusted financial institutions across many regions. Lastly, there is a pervasive low or no trust in Facebook's ability to protect young users, especially among those with higher educational attainment.

- Trust in grocery stores to ensure food safety tends to decrease with age, with younger respondents showing higher levels of complete trust [\[1\]](#)
- 42.86% of respondents in Wyoming express "A lot" or "Somewhat" trust in major companies [\[2\]](#)

Type a message and press enter



# (4) AI as a Briefer – research/risks/opportunity

- **Natural Language Querying:** Develop and evaluate systems that allow users to interact with survey datasets through conversational interfaces, including testing accuracy, usability, and scalability for diverse user groups.
- **Integration of External Data:** Explore methods for combining survey data with external datasets (e.g., administrative records, social media, economic indicators) to enhance validation and enable richer cross-survey comparisons.
- **Guardrails:** Investigate techniques to prevent AI-driven reporting tools from surfacing spurious correlations or oversimplified narratives.
- **Improving Data Literacy:** Design interventions that help non-specialists critically interpret survey output.
- **Dynamic and Multilingual Reporting Systems:** Assess the feasibility and accuracy of real-time translation and adaptive reporting formats to make survey insights globally accessible and culturally relevant.

# transparency and standards - principles

- **What do consumers need to know** to understand the potential bias and limitations of the work
- **What do researchers need to know** to reproduce the survey
- Disclosure should be **organized by task**
- Disclosure should have a **complete checklist**, per task, any given venue (or instance) can choose to leave aspects blank

# transparency and standards - tasks

**(1) AI as the Colleague**

**(2a) AI as the Interviewer**

**(2b) AI as the Respondent**

**(3a) AI as the Transcriber and Translator**

**(3b) AI as the Data Cleaner**

**(3c) AI as the Labeler**

**(3d) AI as the Modeler**

**(4) AI as the Briefer**

**(5) AI as the Workflow**

# transparency and standards – checklist

## 1. Task Performed by AI

- **Purpose**
- **Description:** Briefly describe what the AI did.

## 2. Human Oversight or Validation

- **Validation Steps:**
- **Validation Details:** How was oversight conducted?

## 3. Model Details

- **Model Name & Version:** The specific AI system used (e.g., *GPT-4.0, GPT-5, etc.*)
- **Model Type:** Whether the AI is **open-source** (publicly available) or **proprietary** (owned by a company).
- **Dates of Access/Use:** When the AI was used for the task (e.g., *November 18, 2025*)
- **Fine-Tuning Status:** Was the model fine-tuned (i.e., adjusting the AI to specialize in a certain task) for survey-related work? (Yes/No).
- **\*Fine-Tuning Details:** If yes to fine-tuning, what data was used, and from what source(s)?
- **\*Source URL:** Link to official model documentation.
- **\*RAG Usage:** if and how RAG was used to ground information, and if so, what documents were used.
- **\*Custom Configurations:**

## 4. Access/Tooling Details

- **Method of Access:** How the model was accessed (e.g., *API, website, embedded in platform*)
- **Instrument or Interface (if applicable):** Where and how the AI was embedded or interacted with (e.g., *Qualtrics integration, custom dashboard, interviewer bot*)

## 5. Core Prompts or Instructions:

- **Representative Prompts:** While exact prompts are preferred, researchers can report high-level plausibly abstracted prompts used to guide the model.
- **\*Exact Prompts:** exact prompts used to guide the model.
- **\*System-Wide Instructions:** Any global settings or instructions guiding AI behavior.

## 6. Additional Enhanced Disclosures:

- **\*Code:** Any scripts or code used to call the AI.
- **\*AI as Interviewer:** Characterization of variance in questions asked. Representative samples of conversations.
- **\*Memory:** Was the system stateful or stateless during interaction? (i.e., does it remember previous interactions and let them affect future interactions).
- **\*Known Biases:** Document any other known biases that could affect results.
- **\*Justification:** Why the model and access/tooling was chosen?
- **\*Failure modes:** If and where the AI failed/was wrong, if they had to override it manually.

## 7. Human Respondents Disclosure

- **Number of Human Respondents:** Report the total number of humans who performed the task (e.g., responded to the survey, validated AI outputs, coded data).
- **Synthetic Data:** If human data was augmented in any way using synthetic data, describe how the synthetic was created and used
- **Do NOT Report:** The number of AI instances used for any task.

# AAPOR taskforce

- Background on AI
- AI's use in Surveys
- Evaluation Framework
- Transparency and Reporting for researchers
- Transparency and Reporting for infrastructure and audiences
- Responsibility to Participants

# transparency and standards - discussion

- What does **human validation** mean?
- How do we accommodate **massive variation in understanding** for both researchers and consumers?
- How do you create **lasting rules** amid constant flux? (i.e., we know these tasks are not constant)
- How do we triage tradeoffs between **ideal and practical**? (e.g., we know that some practitioners consider their prompts proprietary)

# transparency and standards - discussion

- What does **human validation** mean?
- How do we accommodate **massive variation in understanding** for both researchers and consumers?
- How do you create **lasting rules** amid constant flux? (i.e., we know these tasks are not constant)
- How do we triage tradeoffs between **ideal and practical**? (e.g., we know that some practitioners consider their prompts proprietary)
- **What is a survey?**

# disruption

- ***Disruption is reimagining workflows to fully leverage AI's capabilities:*** not augmenting current workflows
- ***AI's advance is inevitable, but its societal trajectory is not:*** imperative for thoughtful individuals and institutions to actively shape its trajectory toward openness, innovation, and broad-based welfare gains
- Surveys can become so much more powerful, more central: scalable, qualitative-enhanced, linked-panels, but need:
- Survey researchers to force transparency, standards, collaboration on models and tools

**Augmentation not Replacement => Surveys flourish and Survey researcher flourish**