National Internet Observatory

Building a shared infrastructure to study the internet

Alexi Quintana Mathé

PhD Student, Lazer Lab

Northeastern University

What’s the future of internet research?

Rolling disaster because of shut down of Twitter (now X) data

Percolating over 2023. Key events:

- Shut down of Twitter’s academic API
- Shut down of other APIs
- Shut down of CrowdTangle, URL data shares from Meta
- Elimination of affordable access to decahose data (10% sample of Twitter)

Political science research on the internet

Case study of social sciences more generally

Looked at top 3 journals in US political science (APSR, AJPS, JOP)

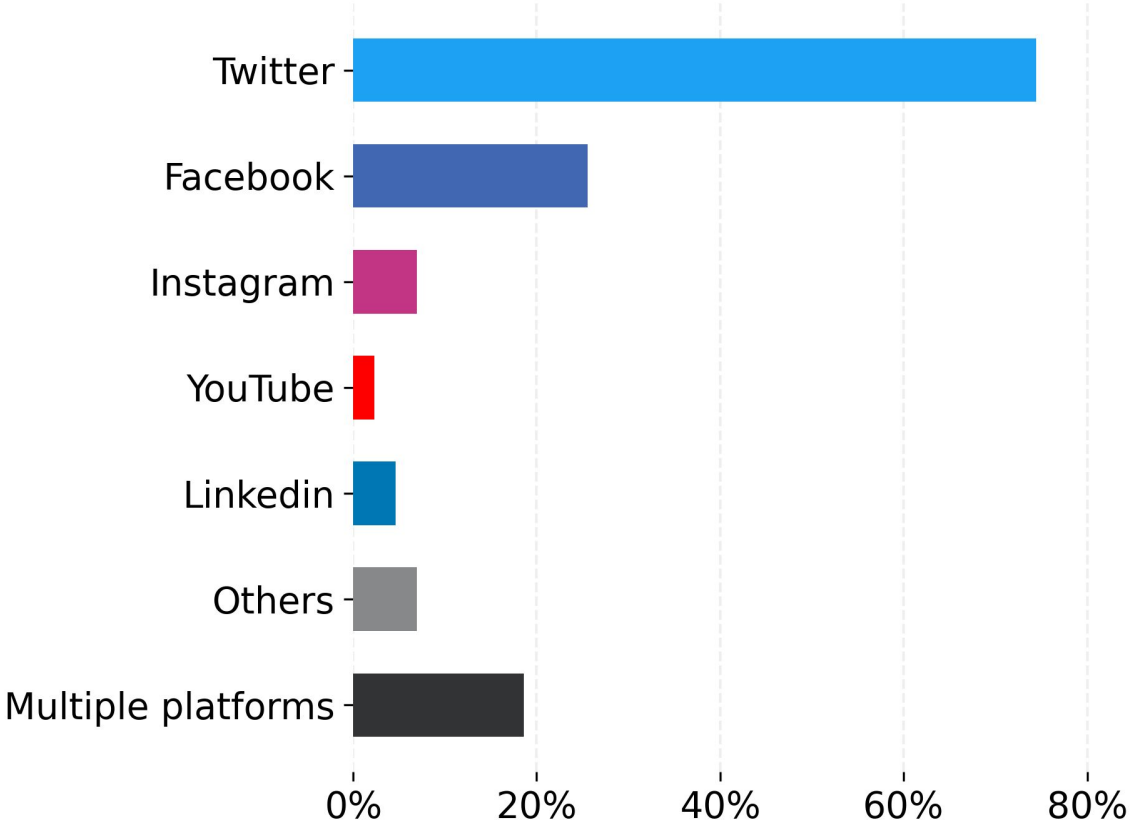
What are the trends over time?

What Internet platforms were focused on?

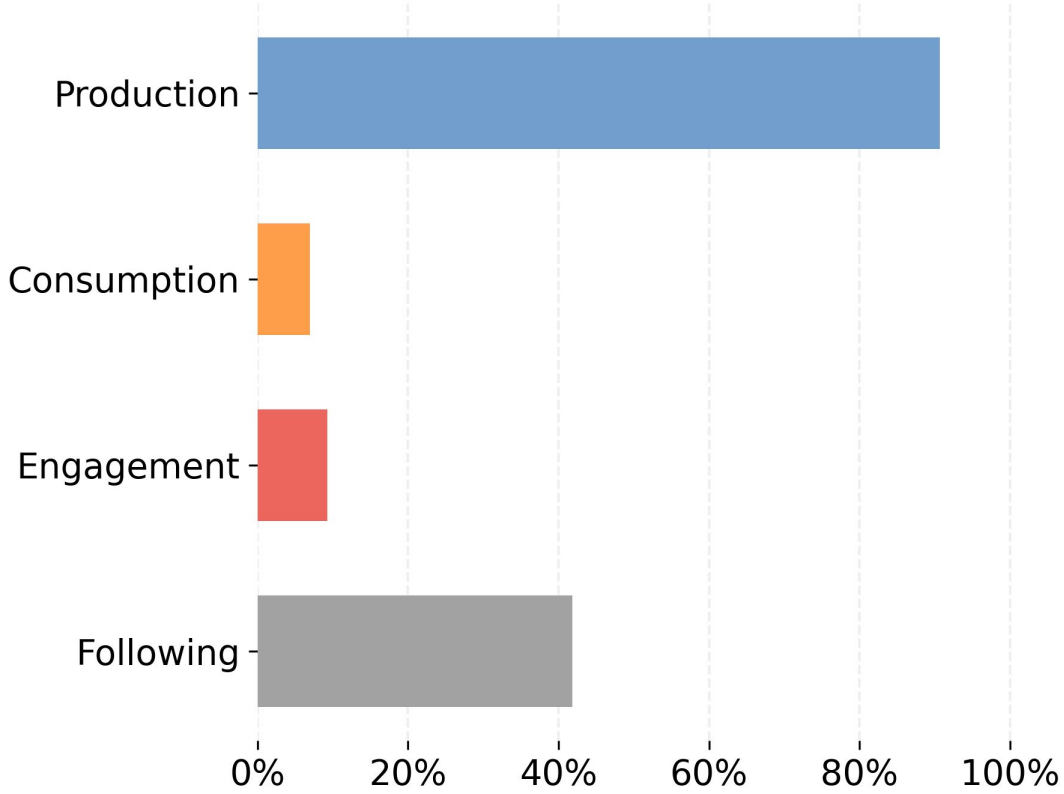
What element of the online experience was examined?

Credit to Kaicheng Yang, Pranav Goel, Mel Allen

Twitter is the dominant source of data



What was studied in the literature?



And how does that match what people *do* on platforms?

Using large scale (~25+k), non-prob- survey (COVID States Project)

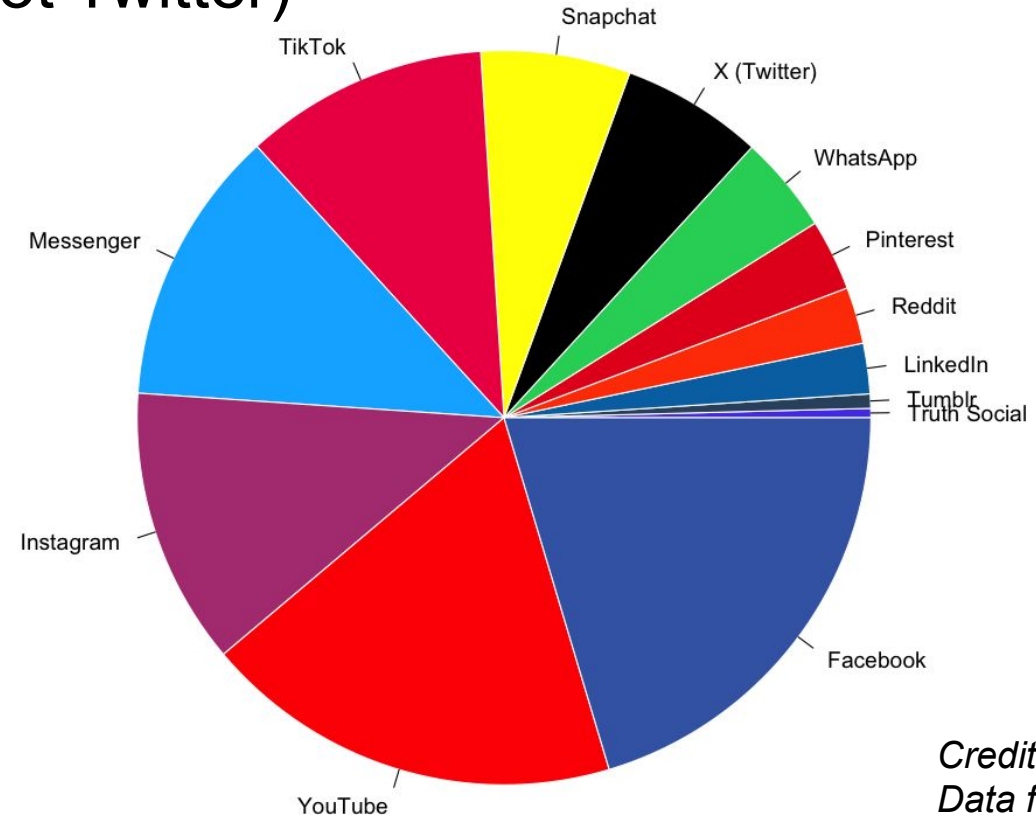
Asked what platforms people used

How often they used them

And how often they posted

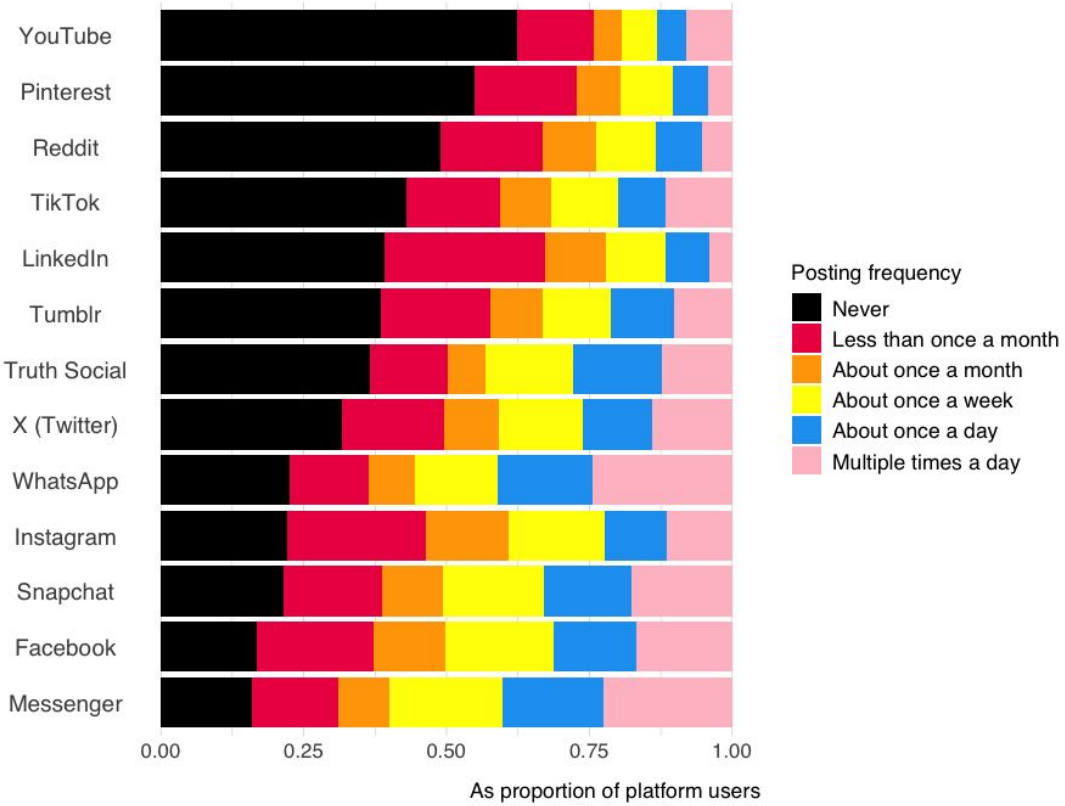
(Omitted platforms < 2% of users)

On what social media do people spend their time on the internet? (not Twitter)



Credit to Ata Uslu
Data from COVID States Project

Most people do not post much (or at all)



Have we been focused on the “right” things?

We have been studying important, interesting things.

But: we’ve only been looking at a small slice of a small slice of what people do on social media (and the internet more generally)

And: we study the “attention economy” while rarely measuring attention.

So: not really.

Possible models…

Leveraging bespoke data collection

Collaborate with companies

Encourage data sharing by companies through policy

Build a shared infrastructure

National Internet Observatory

The NIO model: Volunteer-sourced data with shared infrastructure

PIs: David Lazer, Christo Wilson and Dave Choffnes
Supported by the NSF #2131929

Northeastern University
Network Science Institute

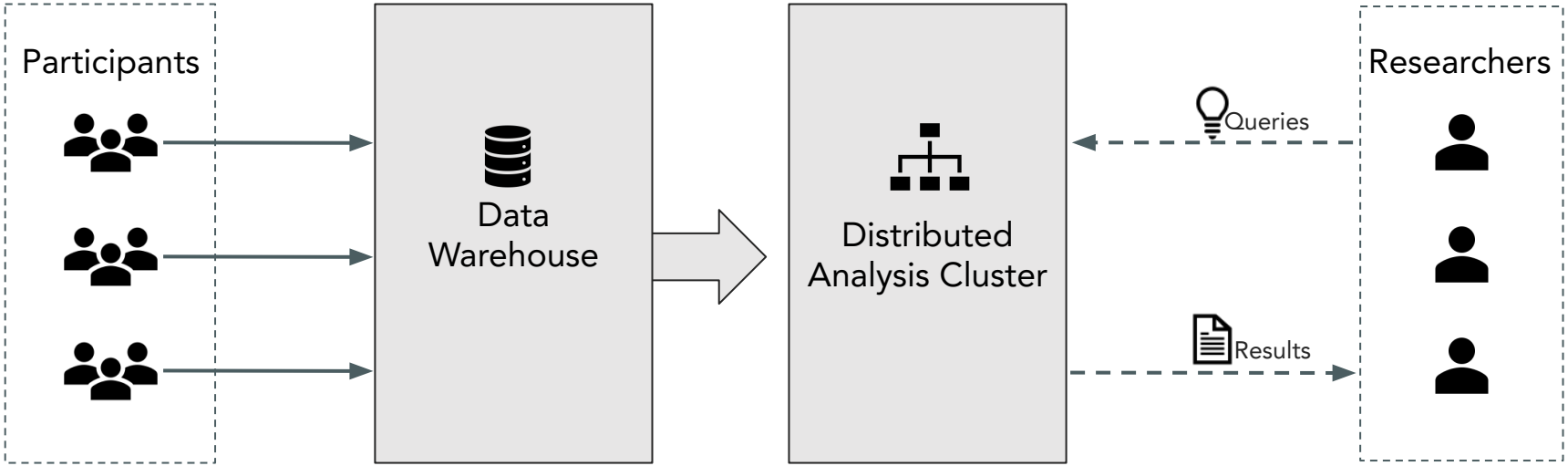
Project Objective

Collect data from a large number of consented participants

Capture various online and survey data of participants (what they see!) and stores the data and information on secure servers.

Provide analytic access to a wide set of academic researchers within a secure, privacy-preserving framework.

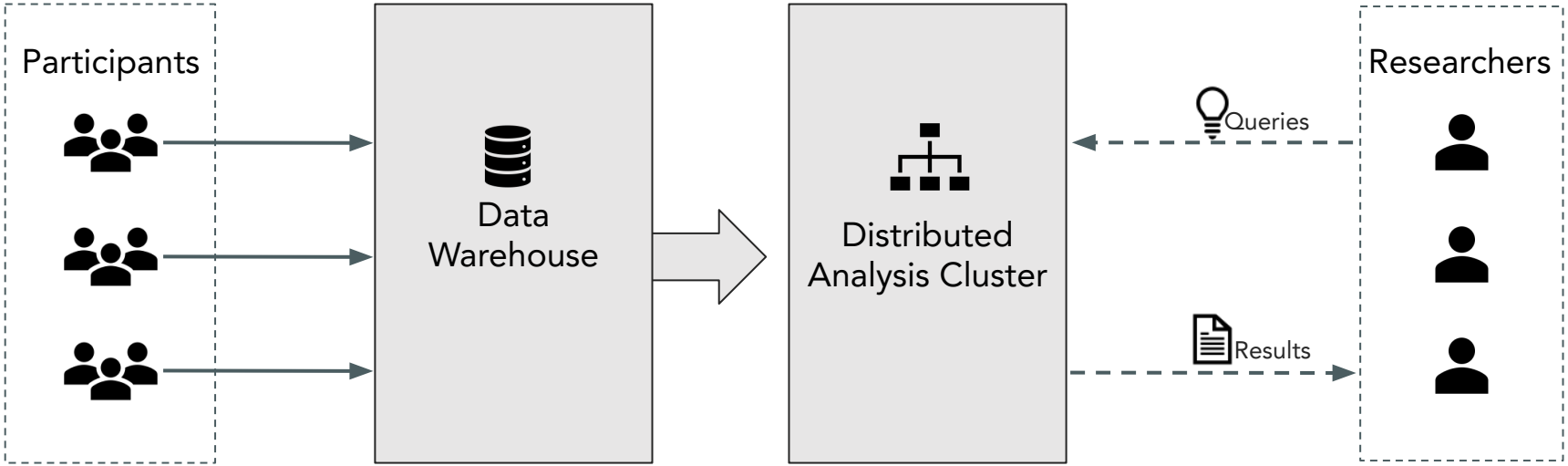
Approach


Participants
Data Warehouse
Distributed Analysis Cluster
Queries
Results
Researchers

Recruitment

Data

Researcher access and ethics



Participant Recruitment

Recruitment pipeline

Online Ads

BOVITZ

VERASIGHT

Recruitment pipeline

Online Ads

BOVITZ

VERASIGHT

Browser extension

Android app

iOS app

Recruitment pipeline

Online Ads

BOVITZ

VERASIGHT

Browser extension

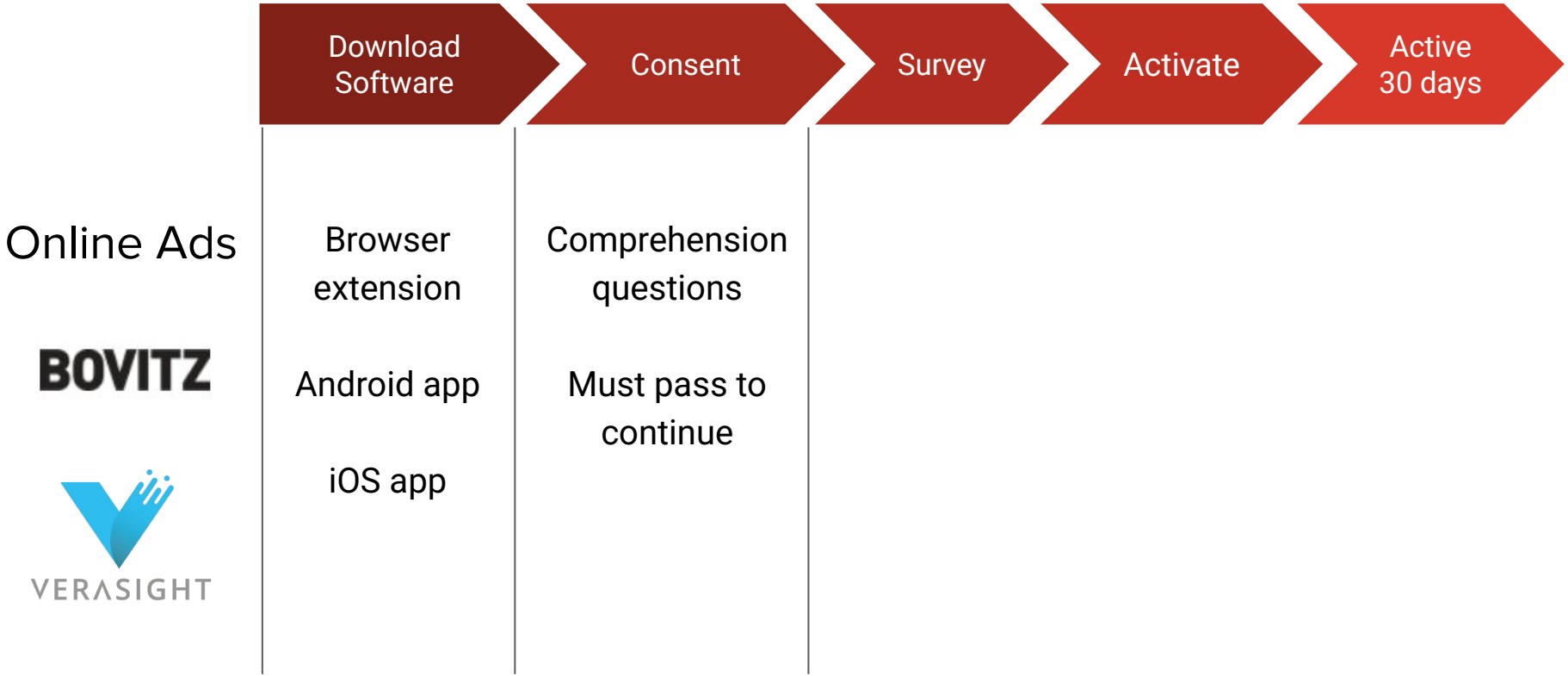
Android app

iOS app

Comprehension questions

Must pass to continue

Recruitment pipeline



Recruitment pipeline

Online Ads

86%

56%

$15-$23 per active 30 days + $10 incentive

BOVITZ

79%

58%

VERASIGHT

78%

56%

$55 per active 30 days + $10 incentive

Recruitment pipeline

Online Ads

86%

56%

$15-$23 per active 30 days + $10 incentive

BOVITZ

79%

58%

VERASIGHT

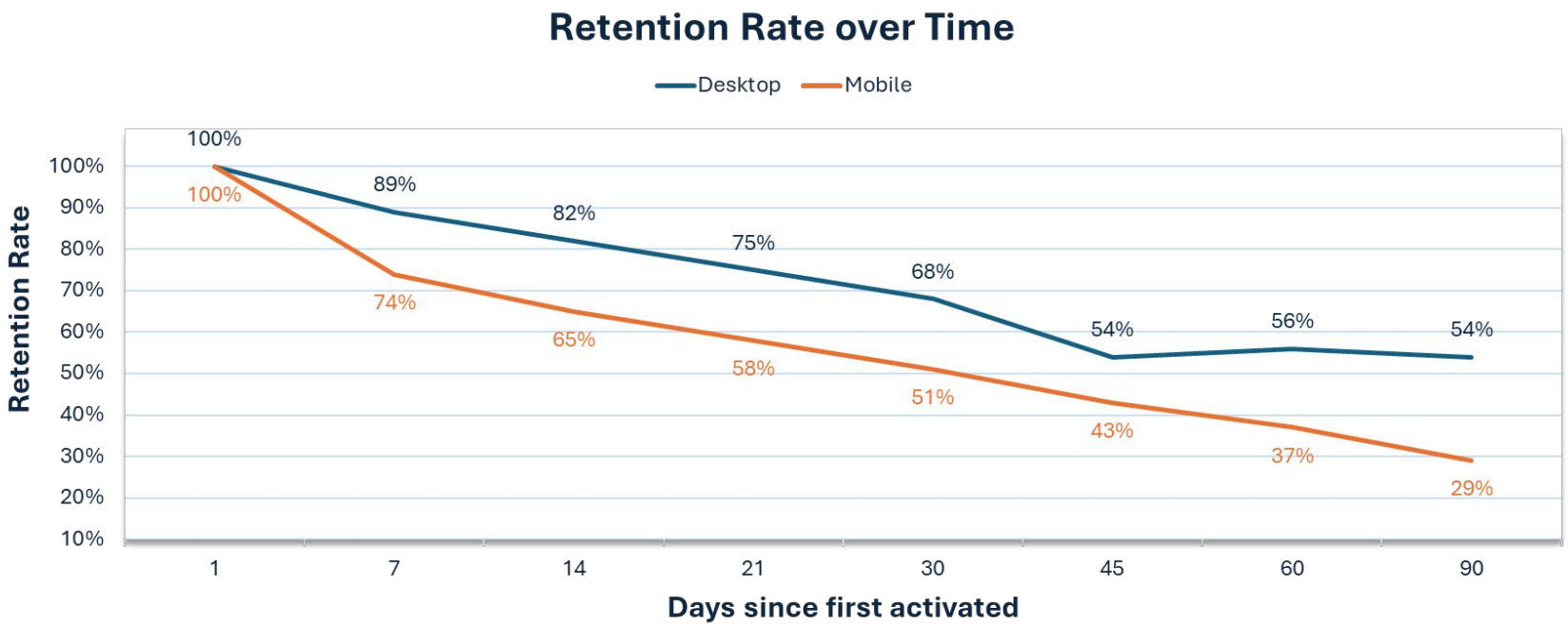
78%

56%

$55 per active 30 days + $10 incentive

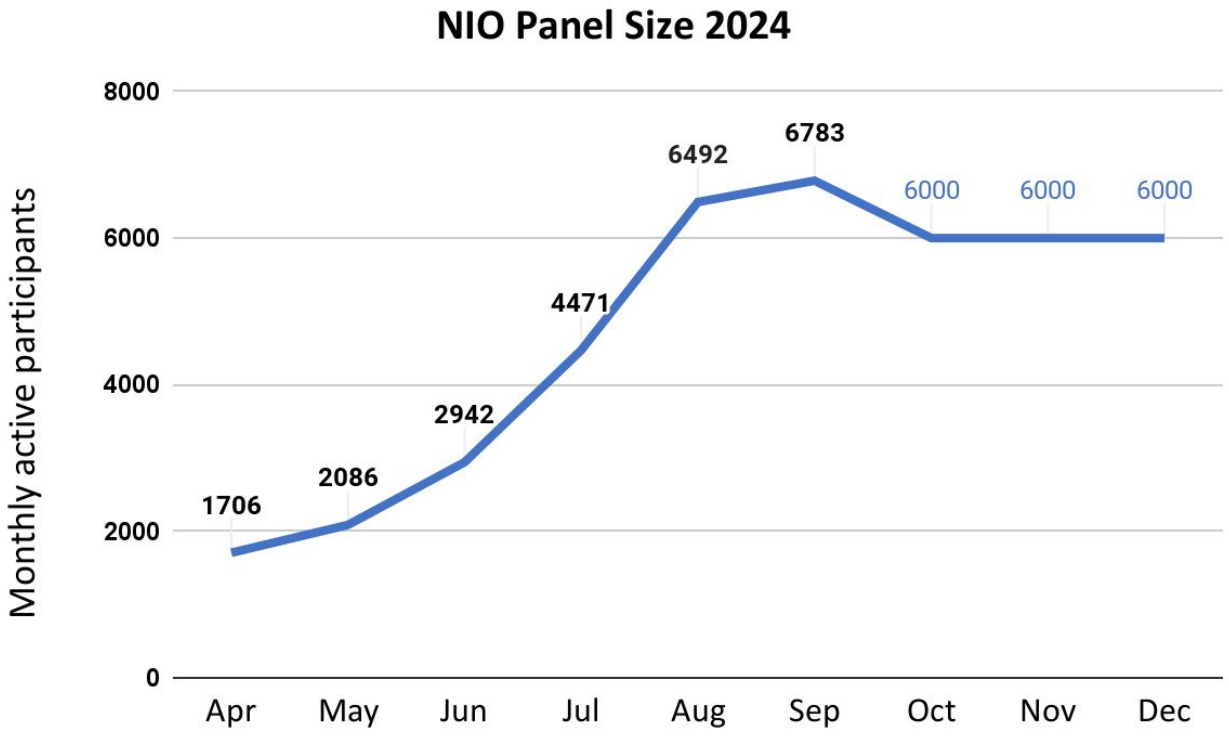
Average monthly recruitment cost was $35, average cost per participant ~$50

Retention Rate



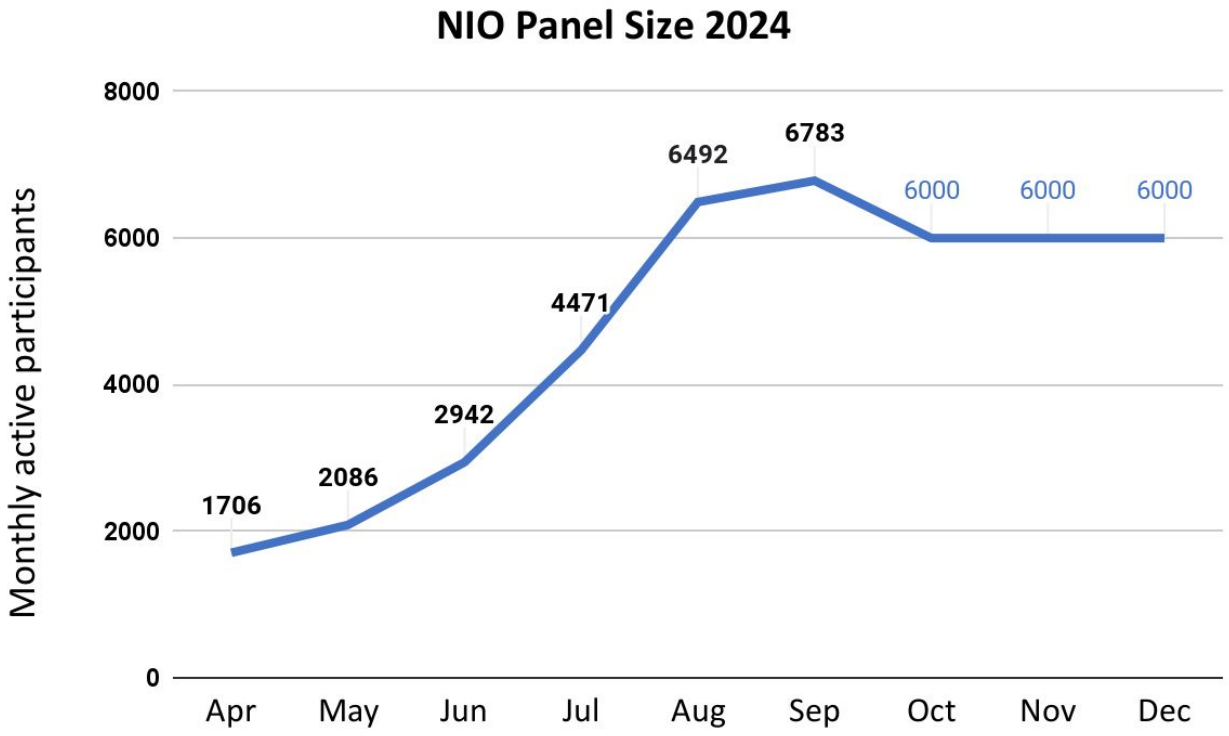
Average monthly attrition rate is roughly 21%

Sample size


NIO Panel Size 2024
Monthly active participants
8000
6000
4000
2000
0
1706
2086
2942
4471
6492
6783
6000
6000
6000
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Sample size

A relatively large sample size is prioritized



Representativity

- We undersample participants without High School, Hispanics, and older adults, and our sample skews liberal compared to National benchmarks.

- Large number of professional survey takers.

Evaluating sample quality: 3 strategies

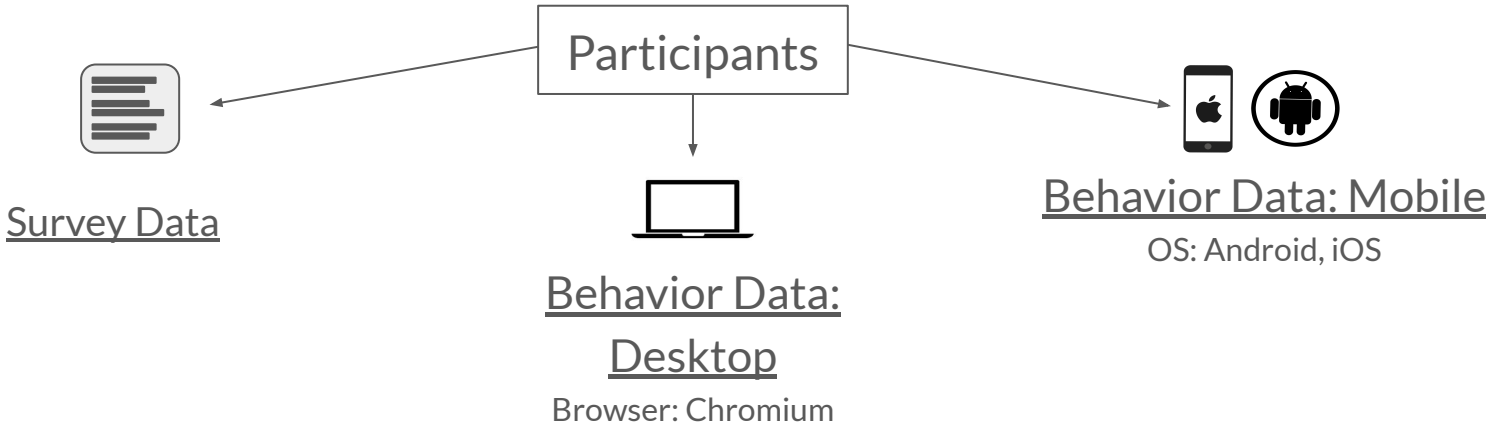
Recruitment of probability sample (tier 1): expensive, but potential partnership with NORC next year.

Survey benchmarks relative to concurrent GSS and ANES surveys, potential re-weighting.

Compare to Nielsen & Comscore browsing data

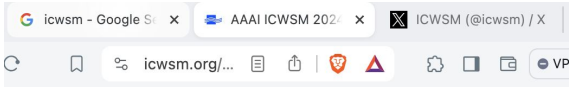
Data Collection

Data collected by NIO



Desktop data

Browsing Activity



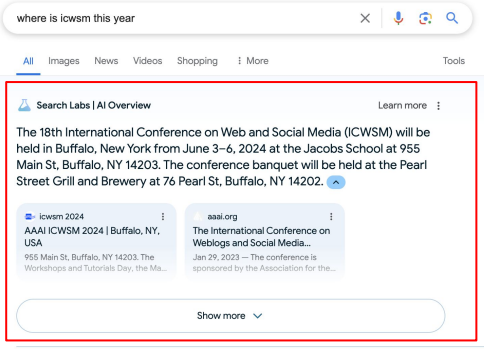
URL sequence

Tab/window transitions

Page navigation

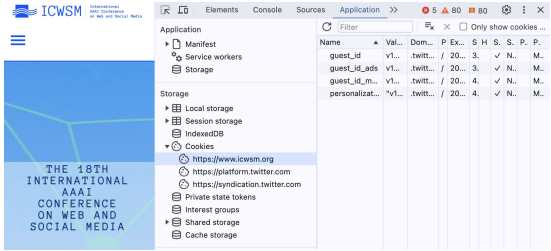
User attention

HTML Snapshots



Google, Bing, YouTube, Amazon, Twitter, Chatgpt, etc.

Browser State

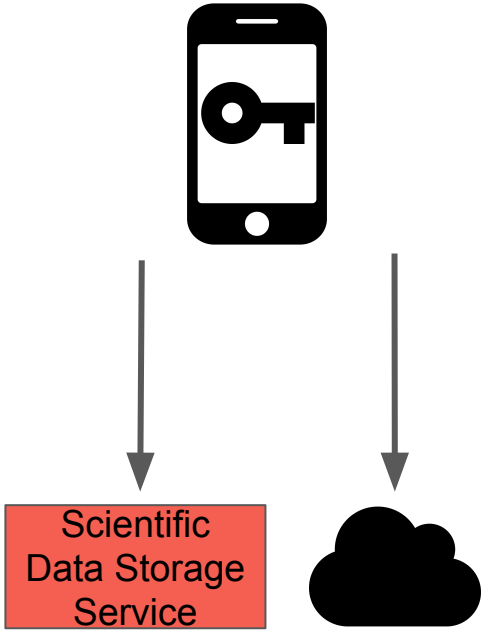


Browser cookies

Google's privacy initiatives

Mobile data

Network Data



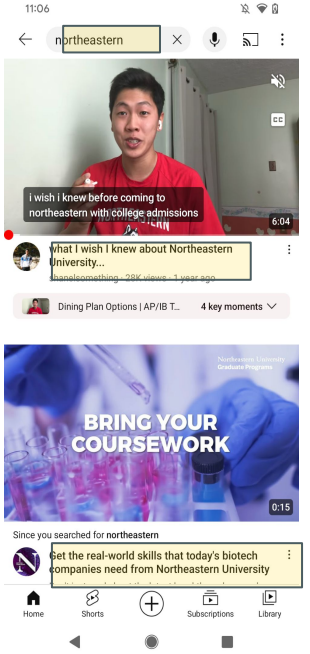
App Usage Data

App name

Launch count

Time in foreground

App content



Surveys

- Initial demographic survey after consent

- Large surveys – every other month, one month in the field
  - 15 minutes each and $3 per complete

- Smaller “micro” surveys
  - Every 1-2 weeks
  - A few minutes and $1 per complete

Larger Surveys

Survey Name	**Dates**	**N**	**# of Participants who Completed Previous Survey**
Mental Health Status (1)	December 2023 – March 2024	1124	-
Politics (2)	April 2024	1050	53*
Psychology (3)	June 2024	1282	531
Politics 2 (4)	August 2024	4401	824
Pre/post election validation (5, 6)	October 2024 & November 2024	~7K each	

Larger Surveys

Response rates of around 55%

Survey Name	Dates	N	# of Participants who Completed Previous Survey
Mental Health Status (1)	December 2023 – March 2024	1124	-
Politics (2)	April 2024	1050	53*
Psychology (3)	June 2024	1282	531
Politics 2 (4)	August 2024	4401	824
Pre/post election validation (5, 6)	October 2024 & November 2024	~7K each	

Ethics and researcher access

The Ethics Challenge

- The status quo for ethical regulation, oversight, and guidance is inadequate for the research enabled by NIO.

nature computational science

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature computational science › comment › article

Comment | Published: 27 July 2023

Enhancing the ethics of user-sourced online data collection and sharing

Michelle N. Meyer, John Basl, David Choffnes, Christo Wilson & David M. J. Lazer

Nature Computational Science (2023) | Cite this article

Metrics

Social media and other internet platforms are making it even harder for researchers to investigate their effects on society. One way forward is user-sourced data collection of data to be shared among many researchers, using robust ethics tools to protect the interests of research participants and society.

https://www.nature.com/articles/s43588-023-00490-7

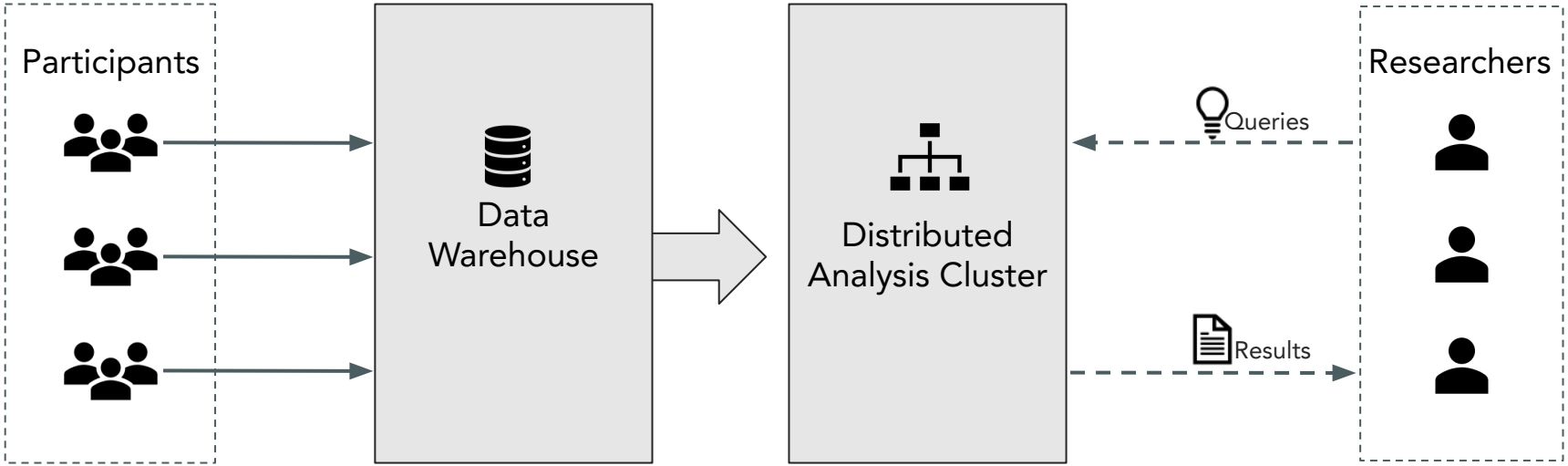
Ethics Toolkit - Overview

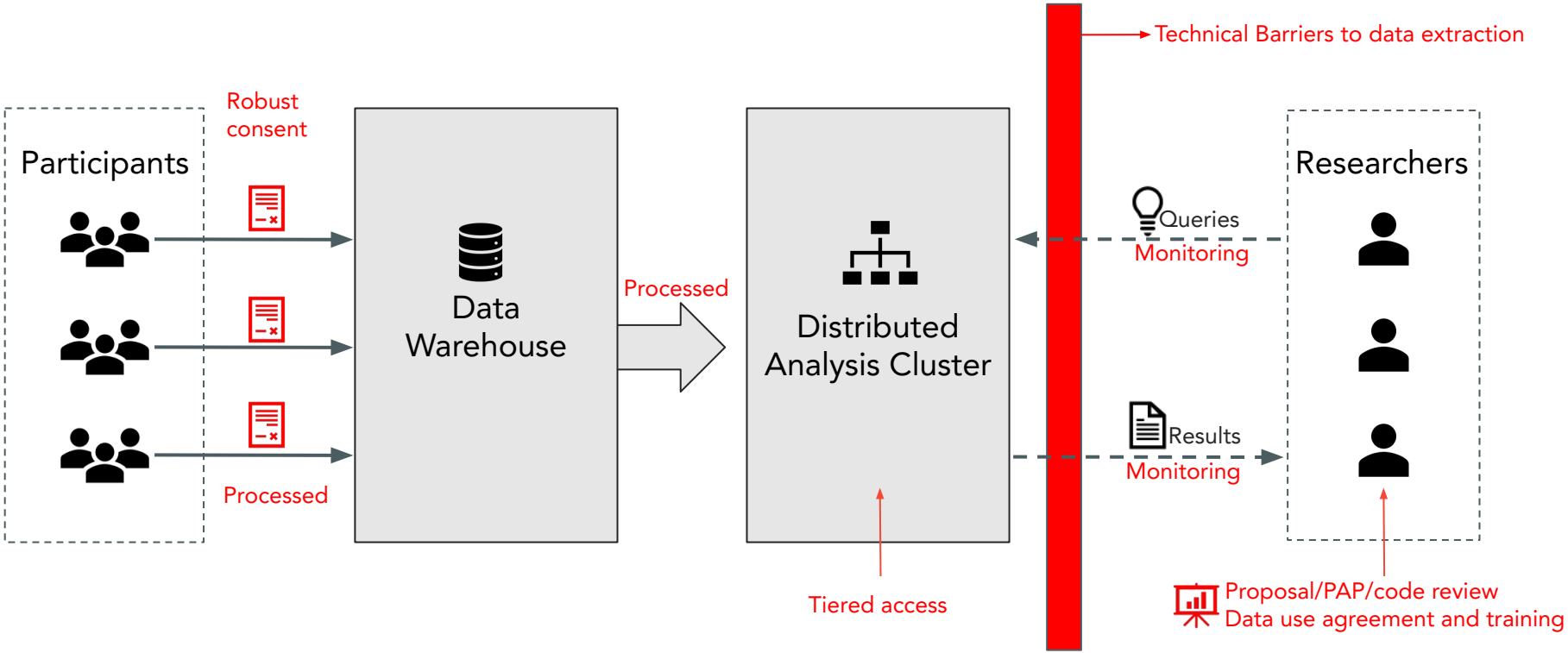
Fig. 1: NIO ethics interventions, mapped to the ends each serves.

Core ends					Examples of NIO ethics interventions
	√				Oversample populations underrepresented in internet and/or social science research
			√		eConsent with teachback questions and a comprehension quiz
√			√		eConsent and NIO website provides instructions about how to manage privacy (e.g., using incognito mode)
			√		Regular participant reminders that they are enrolled in NIO
√	√		√	√	Participant engagement and feedback via surveys, focus groups, and/or a standing community advisory board
	√			√	Researchers have early access to NIO sample characteristics to appropriately gauge feasible research questions
√	√	√			Beta testing data access with small group of trusted faculty colleagues who provide feedback on vulnerabilities
		√			Institutional buy-in: institutional official must sign DUA and inform NIO if a researcher falls out of good standing
√	√	√			Required modular ethics training for researchers
√					Access is provided to individual researchers only and protected by security best practices
√	√	√			Standardized (not project-specific) data use agreement (DUA) countersigned by researcher institution, • No attempted re-identification • No attempt to exfiltrate, publish, or redistribute data • Immediately alert NIO of unexpected data privacy issue • No linking NIO data with outside or other NIO data without express written permission of NIO • Acknowledge consequences for violation (e.g., reporting to IRB/funder; expulsion from NIO)
√	√	√			Researcher application for analytic access to data: • Research question(s) • Description of data requested and justification for each data element • Analytic approach • List of all project personnel who would have access to the data/results • Potential benefits of the research • Description of risks of the project (referring to appropriate training modules) • Assessment of distribution of risks and potential benefits across different groups
√		√			Data remains on NIO servers
√	√				Review of source code, in some cases
√					Data access limited to specific project need
√					Query-return restrictions (e.g., hide results for fewer than *t* participants) in some cases
		√			Real-time passive monitoring of NIO data use to detect data exfiltration
			√	√	Continuously updated list of NIO-based research posted to public study website
√	√	√			Institutional reminders that specific faculty are active NIO users
		√		√	Random and for-cause audits: including manual review of individual monitoring logs, aggregate patterns of data access, and publications

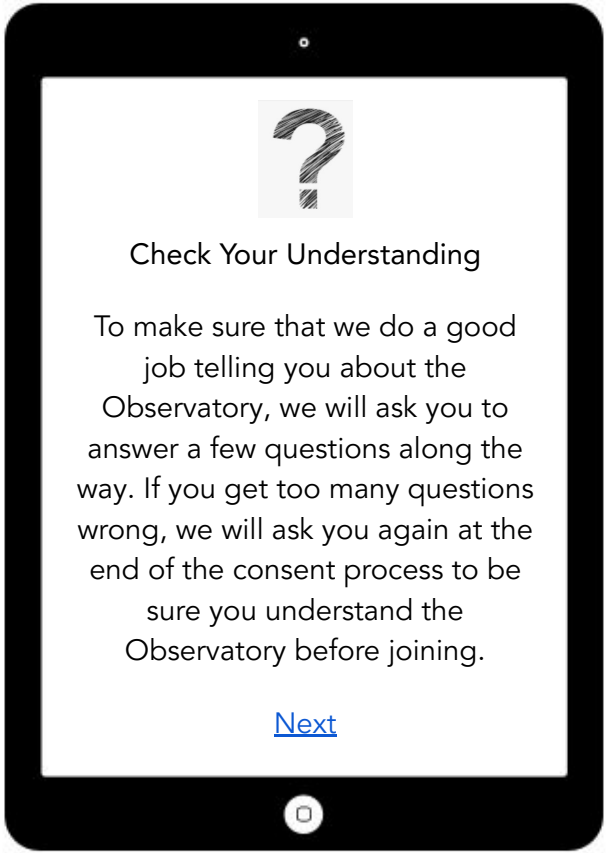
- Minimize privacy risks to participants and bystanders
- Ensure scientifically and socially responsible use of data
- Ensure researcher compliance
- Respect participant autonomy
- Promote transparency

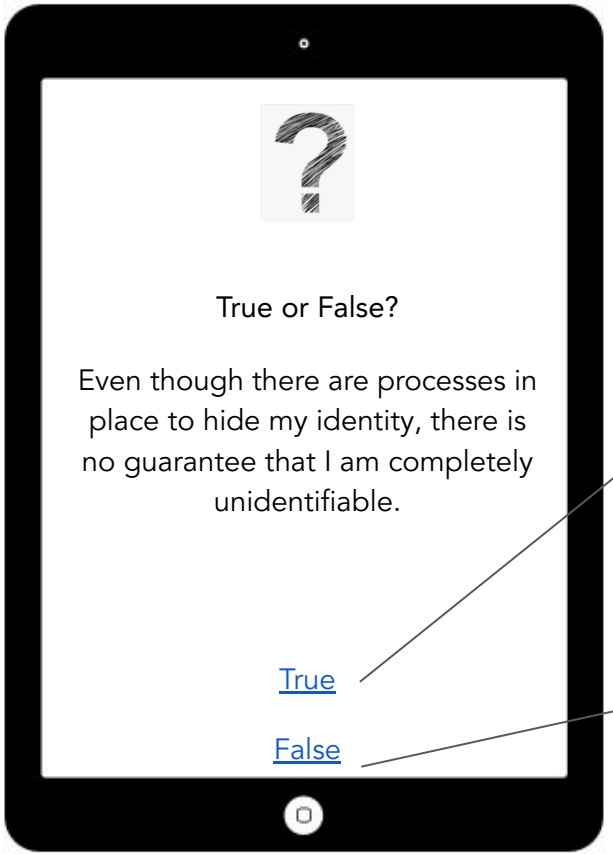
https://www.nature.com/articles/s43588-023-00490-7


Participants
Data Warehouse
Distributed Analysis Cluster
Queries
Results
Researchers


Technical Barriers to data extraction
Participants
Robust consent
Data Warehouse
Processed
Distributed Analysis Cluster
Queries
Monitoring
Researchers
Results
Monitoring
Processed
Tiered access
Proposal/PAP/code review
Data use agreement and training

Toolkit - eConsent “teachback” questions





Those who get <5 of 7 these teachback questions correct must answer the same 7 questions and get at least 5 correct to enroll.

That’s correct: Even though there are processes in place to hide your identity, there is no guarantee that you are completely unidentifiable.

Not quite: Even though there are processes in place to hide your identity, there is no guarantee that you are completely unidentifiable.

Research projects

Case study: DISCO

- DISCO: Depression, Isolation, and Social Connectivity Online

- NIO data with its survey infrastructure combined with online activity data enables the study of various aims in this project, for example: **examine the association between online social activity and depressive symptoms**.

- Won an NIH grant and started in Spring 2023.

Summary of research design

Goal: capture longitudinal, objective measures of online behavior and examine the association between online social behavior and depressive symptoms.

Question: Can other forms of social interaction online mitigate these risks, or might they actually increase depression liability?

Hypothesis: behavioral measures of social media use will associate with magnitude of depressive symptoms, explaining additional variance beyond that captured solely by self-report.

Robertson et al (Nature, 2023)

Article

Users choose to engage with more partisan news than they are exposed to on Google Search

https://doi.org/10.1038/s41586-023-06078-5

Received: 17 February 2022

Accepted: 12 April 2023

Published online: 24 May 2023

Check for updates

Ronald E. Robertson[1,2]✉, Jon Green[2], Damian J. Ruck[2], Katherine Ognyanova[3], Christo Wilson[2,4] & David Lazer[2]

If popular online platforms systematically expose their users to partisan and unreliable news, they could potentially contribute to societal issues such as rising political polarization[1,2]. This concern is central to the 'echo chamber'[3–5] and 'filter bubble'[6,7] debates, which critique the roles that user choice and algorithmic curation play in guiding users to different online information sources[8–10]. These roles can be measured as exposure, defined as the URLs shown to users by online platforms, and engagement, defined as the URLs selected by users. However, owing to the

Robertson et al (Nature, 2023)

Abstract:

If popular online platforms systematically expose their users to partisan and unreliable news, they could potentially contribute to societal issues such as rising political polarization. This concern is central to the ‘echo chamber’ and ‘filter bubble’ debates, which critique the roles that user choice and algorithmic curation play in guiding users to different online information sources. These roles can be measured as exposure, defined as the URLs shown to users by online platforms, and engagement, defined as the URLs selected by users. However, owing to the challenges of obtaining ecologically valid exposure data—what real users were shown during their typical platform use—research in this vein typically relies on engagement data or estimates of hypothetical exposure. Studies involving ecological exposure have therefore been rare, and largely limited to social media platforms, leaving open questions about web search engines. To address these gaps, we conducted a two-wave study pairing surveys with ecologically valid measures of both exposure and engagement on Google Search during the 2018 and 2020 US elections. In both waves, **we found more identity-congruent and unreliable news sources in participants’ engagement choices, both within Google Search and overall, than they were exposed to in their Google Search results. These results indicate that exposure to and engagement with partisan or unreliable news on Google Search are driven not primarily by algorithmic curation but by users’ own choices.**

Highlights for replicating Robertson et al (Nature, 2023) with NIO data

- Findings in Robertson et al 2023 replicate with NIO data – both in terms of patterns and underlying significant testing.

- We replicated this with NIO in a fraction of the time it took to conduct experiments in the original paper: a few weeks instead of multiple years.

- We worked with roughly 3 times the number of users (as of May 2024), and currently, we have at least 10 times the amount of data as used in Robertson et al 2023.

Highlights for replicating Robertson et al (Nature, 2023) with NIO data

- We can feasibly repeat this replication regularly and check if the paper's findings holds over time, which can help monitor changes in Google's search output.

- The nature of our data collection and our other data products allows to extend this study to many other platforms (for example, we can use data from Facebook, Reddit, Twitter, etc.).

  We can study user engagement with websites compared with platform exposure both across platforms and over time: something that was not possible a year ago!

  - We have already begun to study other platforms like Facebook.

Large Language Models Usage

Between 1/1/2024 and 9/14/2024

- **ChatGPT**: 1166 users, 37314 visits, mean 85.72 minutes per day
- **Gemini**: 350 users, 9512 visits, mean 28.04 minutes per day
- **Claude**: 75 users, 2871 visits, mean 6.65 minutes per day
- **Copilot**: 162 users, 623 visits, mean 1.31 minutes per day

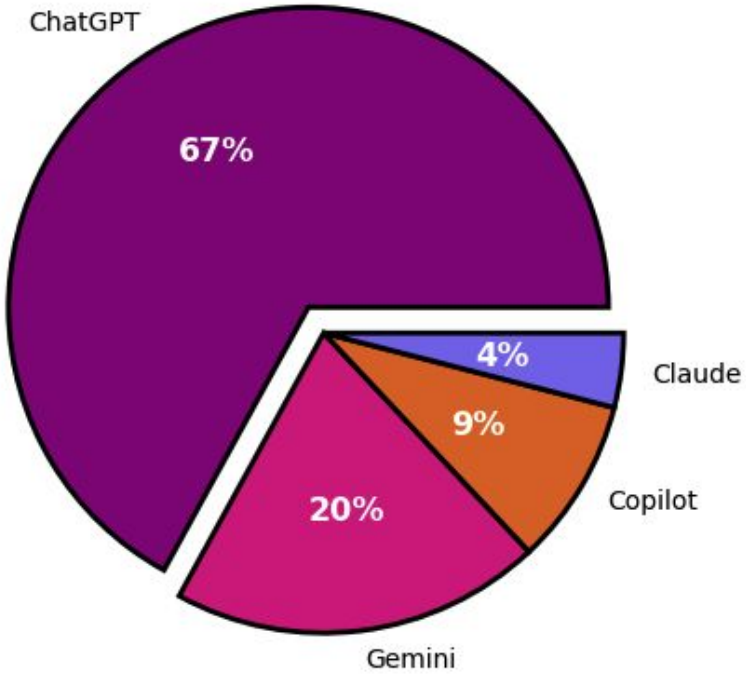
31.4% of the active participant pool** used at least one of these LLMs between Jan 1 and Sep 14, 2024

** those who contributed website visit data for at least 7 different days

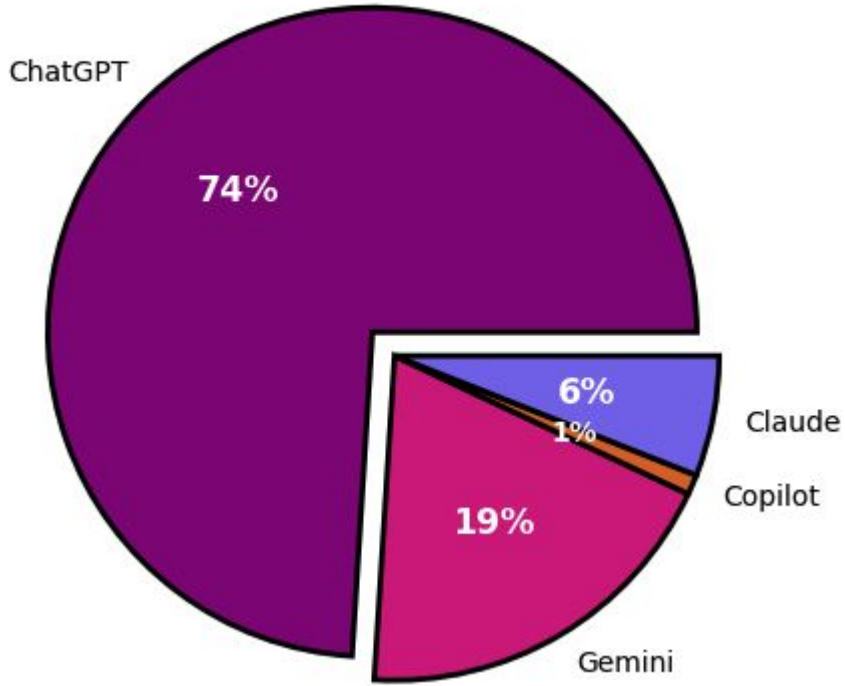
mean time per day is calculated for those days that the LLM was visited

LLM Market Shares

LLM User Market Share (Web Extension)

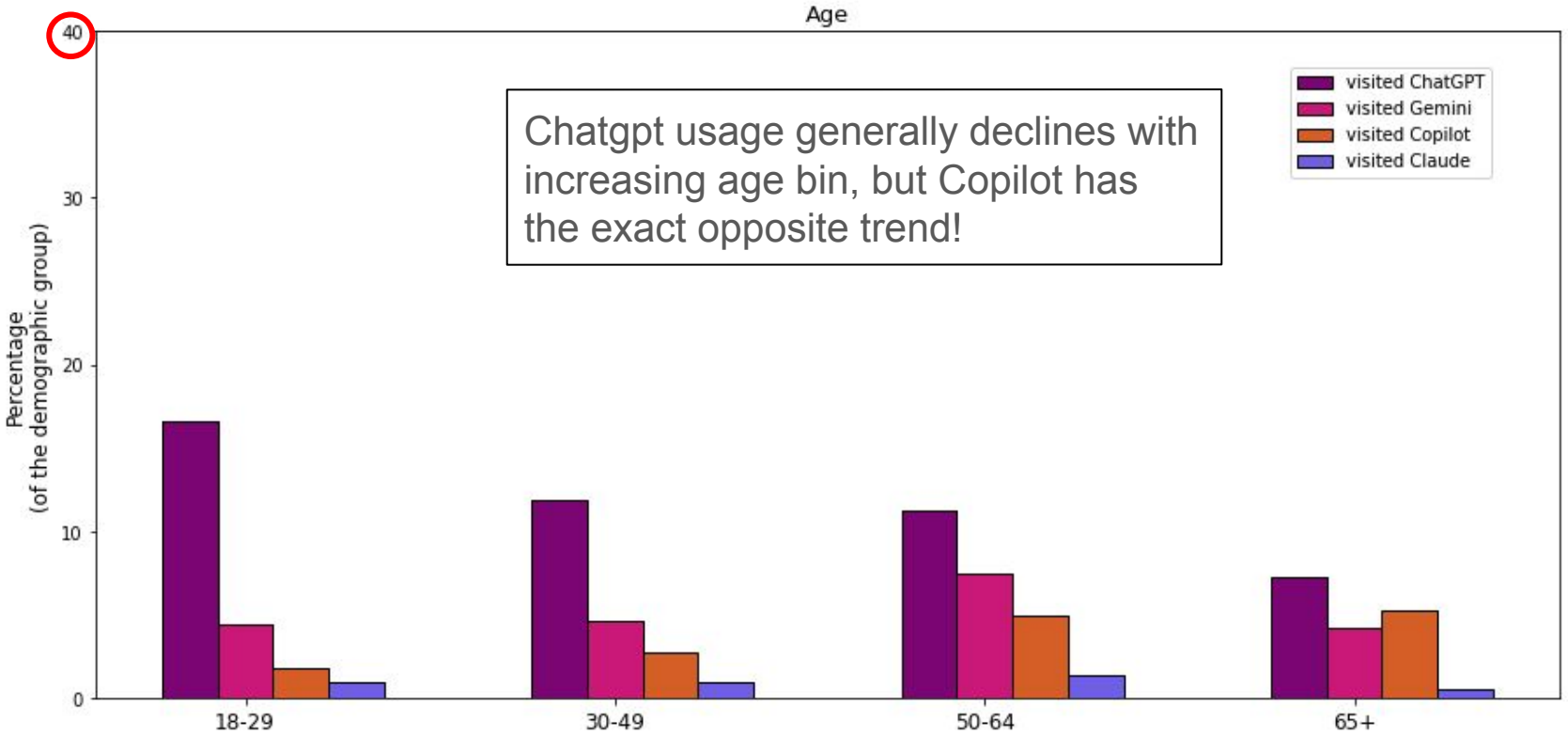


LLM Visits Market Share (Web Extension)



Who uses LLMs? Age distribution for LLM users

ChatGPT vs Gemini vs Copilot vs Claude



Research avenues on LLM usage

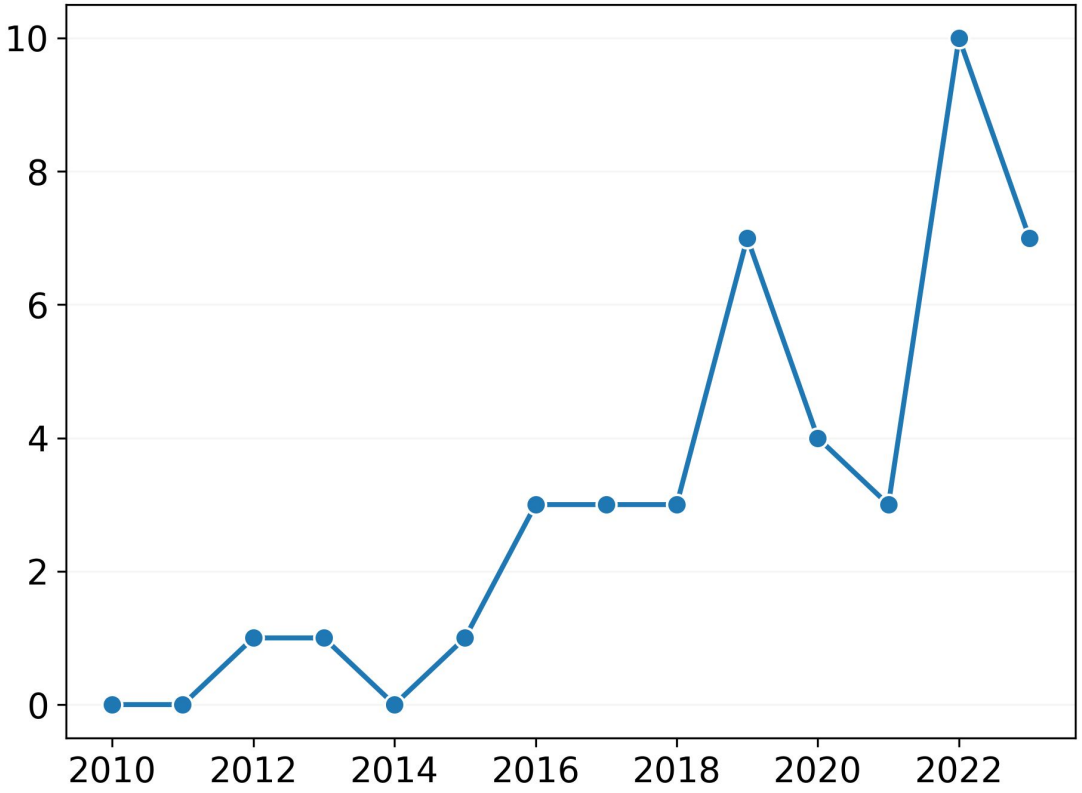
- Paired with the scraped HTLM content on LLM websites, we will be able to investigate how users use LLM’s, and how usage varies for different sociodemographic groups.

- This research would allow answering questions on the impact of these new technologies on society.

Thank you!

Appendix

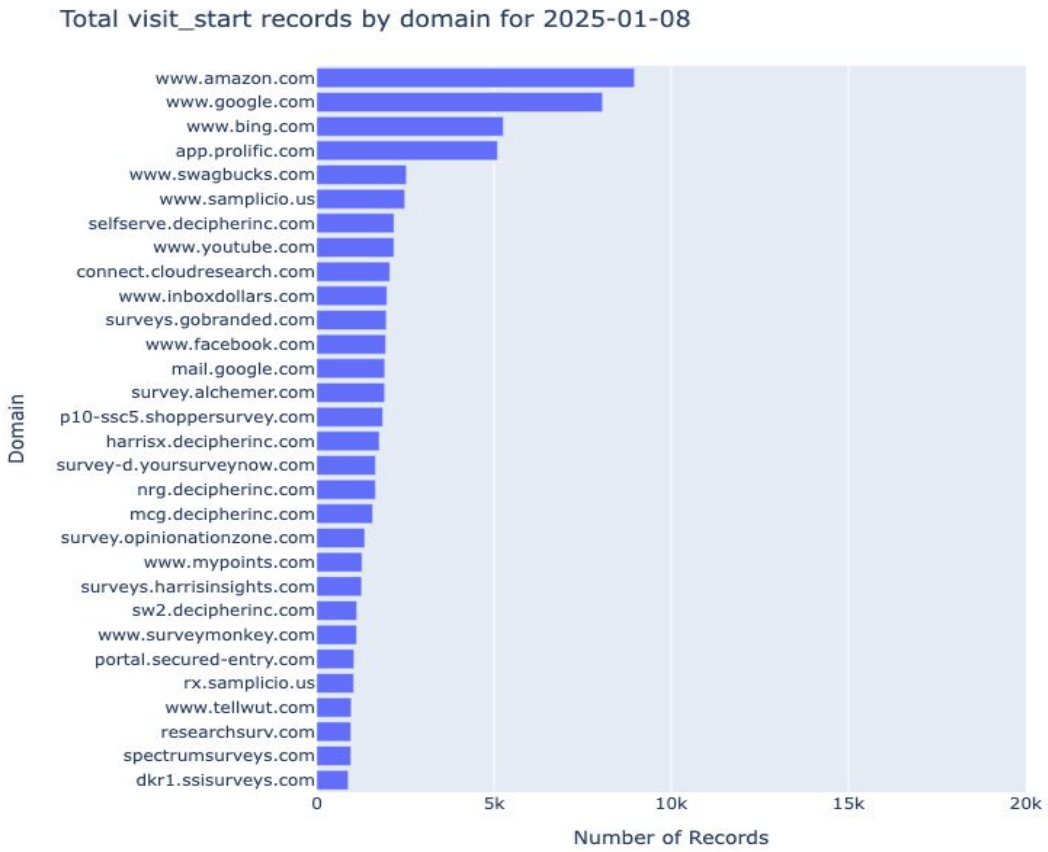
Number of internet papers in top 3 political science journals



2023-24 incomplete

Is the disappearance of Twitter as a data resource a disaster?

Population of professional survey takers



Panel Representation

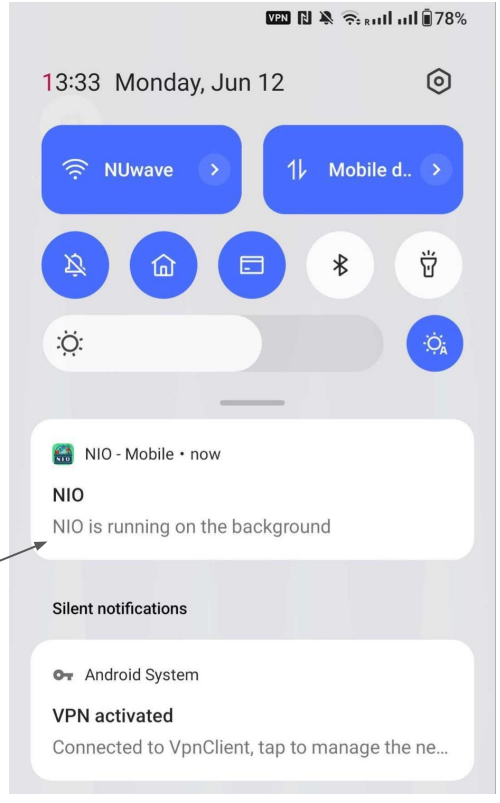
	Active Participants	National Average	Active N
Gender (Women)	48%	51%	3112
Education (HS or less)	17%	38%	1117
Race (White)	70%	75%	4506
Race (Black)	20%	14%	1307
Race (Hispanic)	7%	19%	490
Age 50-64	16%	24%	1036
Age (65+)	7%	22%	463
Income > $100K	29%	34%	1884

Panel Representation

	Active Participants	ANES 2020	Active N
Extremely liberal	13%	5%	865
Liberal	21%	16%	1376
Slightly liberal	12%	12%	793
Moderate	27%	27%	1748
Slightly conservative	11%	12%	676
Conservative	11%	22%	720
Extremely conservative	4%	6%	255

Toolkit - Dynamic Consent

Reminder of enrollment



Transparency about data use (& opportunity to update priors)

Currently there are no Approved Projects, but please check back as this is where we will list all of the approved projects that use this data

Researcher Application Process Flowchart

- Basic research ethics
- Specific issues related to NIO data

Each person who accesses or analyzes NIO data (including summary statistics) must complete ethics training and agree to the CoC

- Research team qualifications
- Analytic & data access justification
- Potential benefits
- Risks
- Risk justification and mitigation

May involve more than one round of review to reach approval

DUA signed by an institutional official at each institution engaged in research

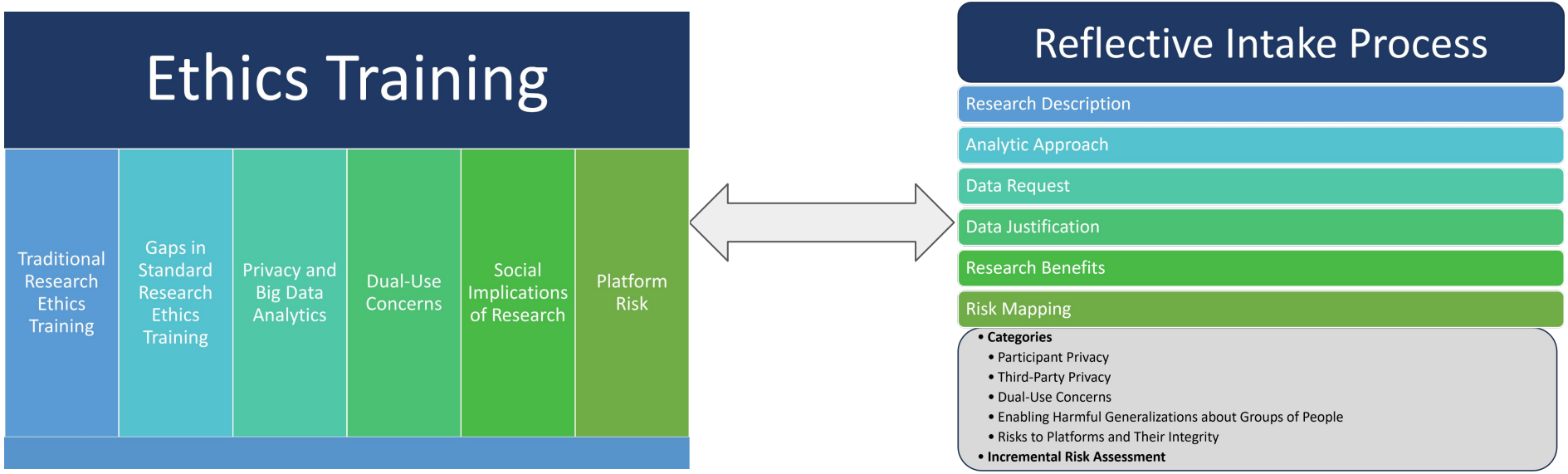
Some other ethical interventions

- Avoid collecting private data.
- Researchers inform when they find privacy issues in the data they work with.
- Export reviews, passive monitoring of NIO data to detect exfiltration, random audits of projects.
- Data remains on NIO servers, restricted queries (no small N queries).
- Participants are provided an option to manage privacy (incognito mode).

Operationalizing Values - Researcher Training

1. CITI SBE courses
2. All of Us training (adapted)
3. Meyer, Basl et al., *Nature Comp Sci* 2023
4. Code of Conduct
5. Zivony et al., *PLOS Comp Bio* 2023

Toolkit - Researcher Training and Reflective Intake



Toolkit - Export Review

- All content (e.g., graphs, tables) reviewed by NIO prior to removal from servers for (1) privacy risk & (2) consistency w/approved research
- Output must be based on minimum unweighted subsample of at least 30
- Relationship between 2+ subsamples must be explained to enable review of privacy risk from disclosure of overlapping sets
- No screenshots permitted

Researcher Application Process Flowchart

- Basic research ethics
- Specific issues related to NIO data

Each person who accesses or analyzes NIO data (including summary statistics) must complete ethics training and agree to the CoC

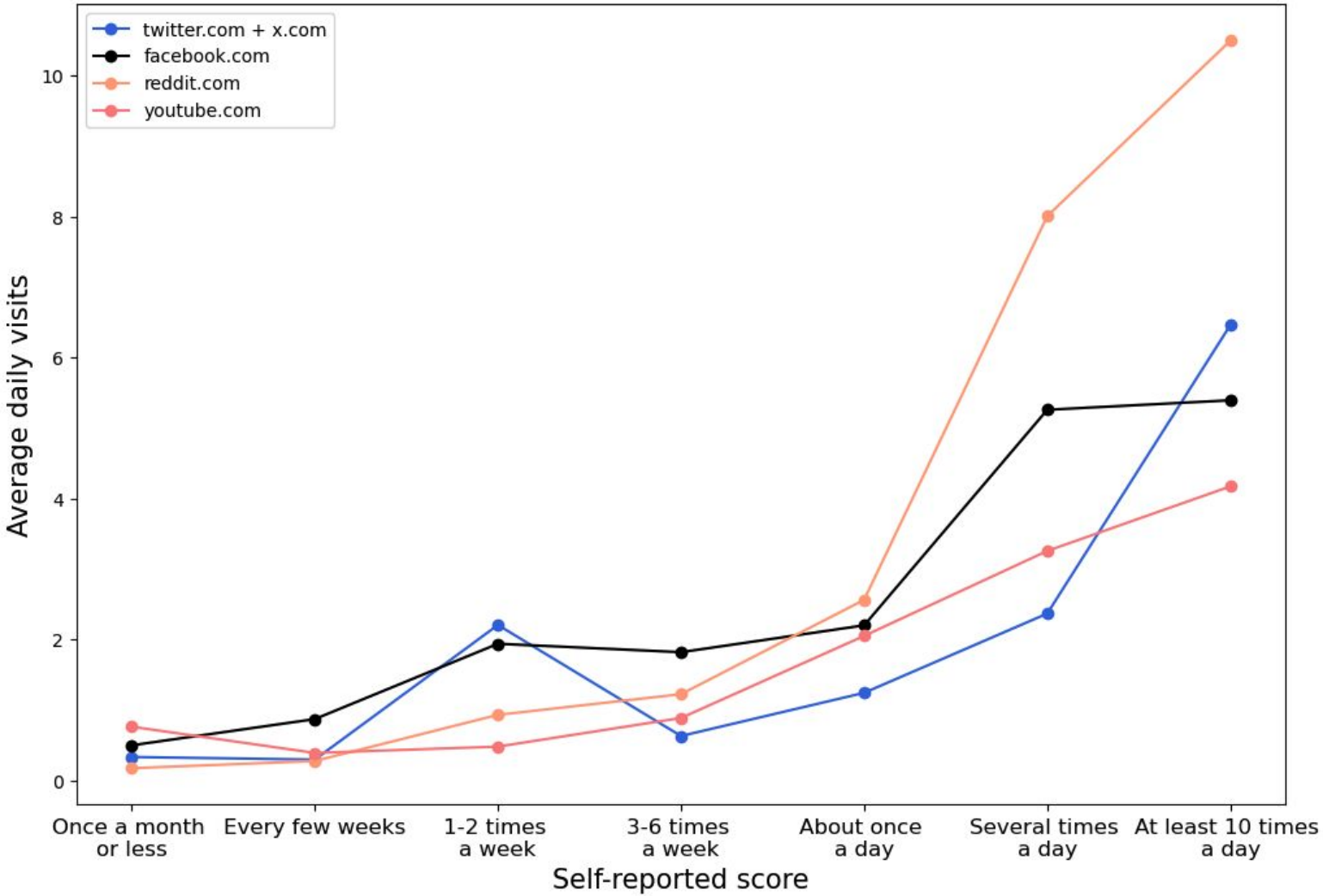
- Research team qualifications
- Analytic & data access justification
- Potential benefits
- Risks
- Risk justification and mitigation

May involve more than one round of review to reach approval

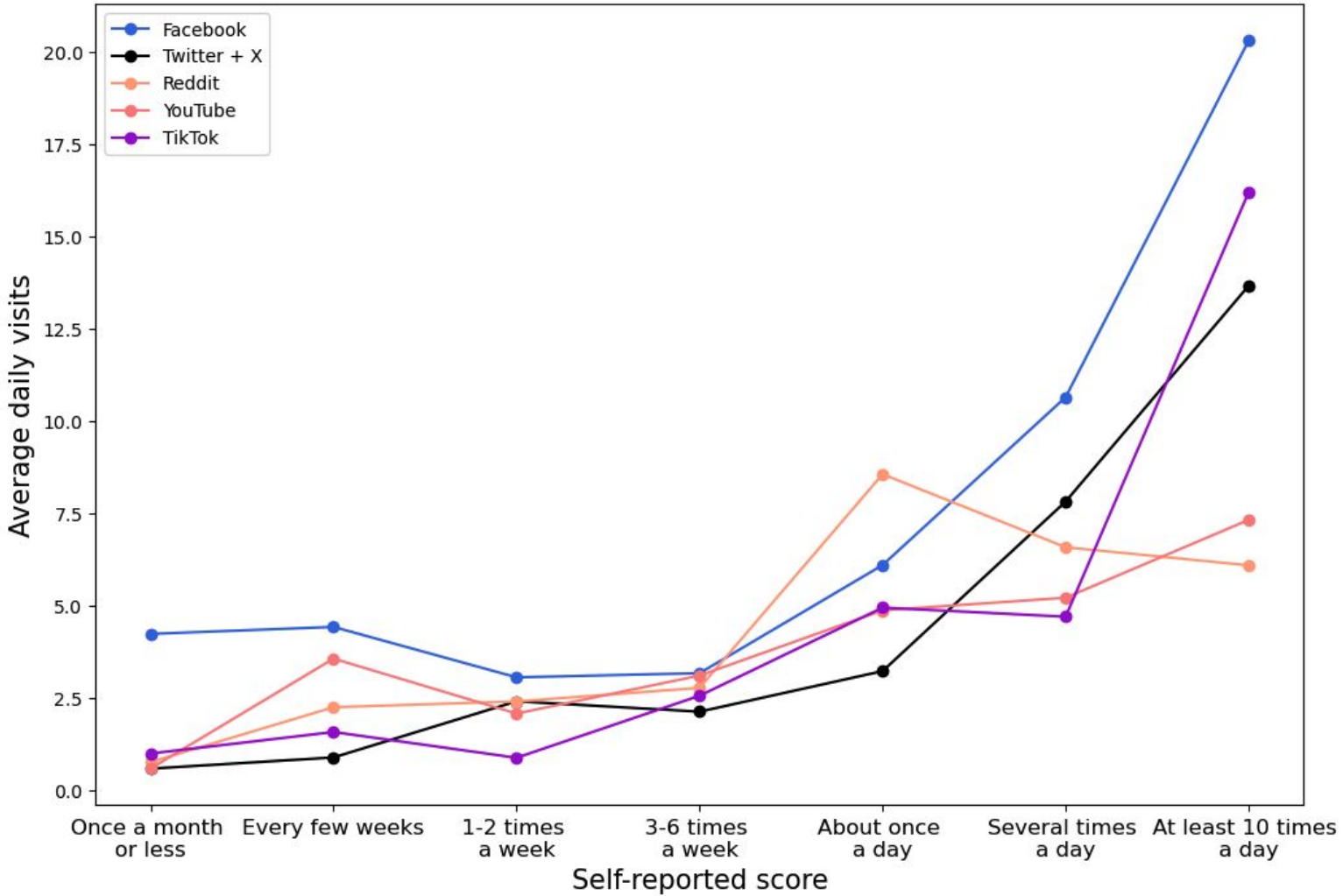
DUA signed by an institutional official at each institution engaged in research

Comparison of self reports of social media usage to observations (desktop & mobile)…

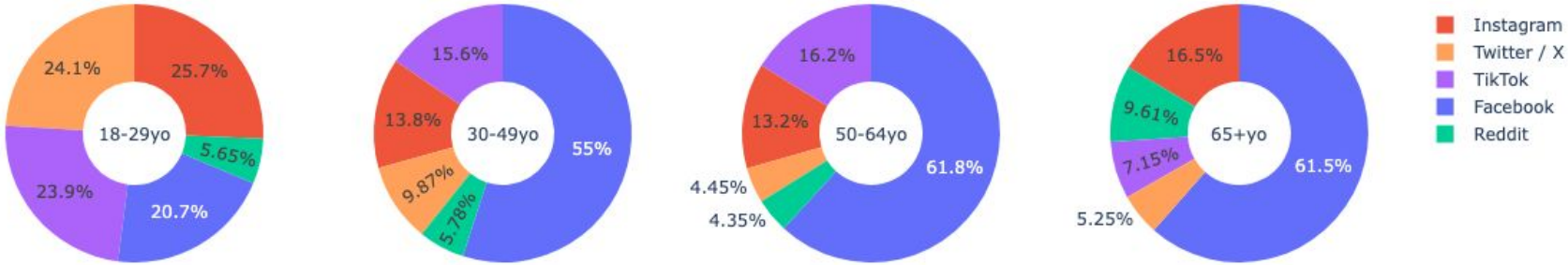
Desktop:


twitter.com + x.com
facebook.com
reddit.com
youtube.com
Average daily visits
0
2
4
6
8
10
Once a month or less
Every few weeks
1-2 times a week
3-6 times a week
About once a day
Several times a day
At least 10 times a day
Self-reported score

Mobile:


Facebook
Twitter + X
Reddit
YouTube
TikTok
Average daily visits
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Once a month or less
Every few weeks
1-2 times a week
3-6 times a week
About once a day
Several times a day
At least 10 times a day
Self-reported score

Social Media App Usage By Age Cohort



Activity is monitored via the app being open every 1 to 15 minutes.

* Filtered by participants active for greater or equal to 3 days

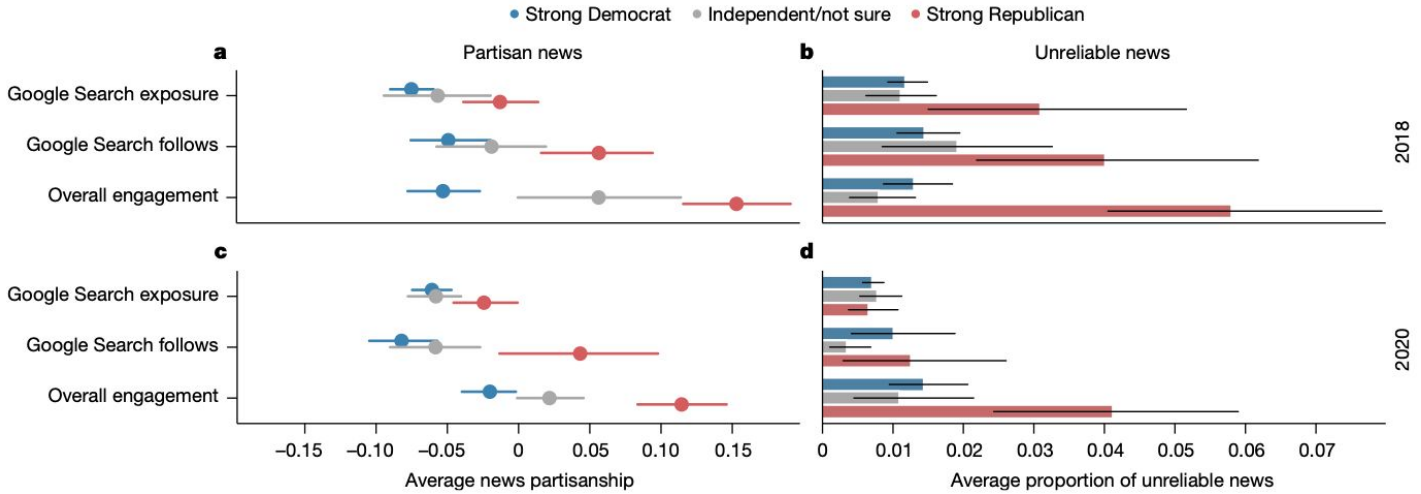
Basic Info

- Robertson et al. – Two Study Waves (2018 & 2020); NIO data comes from May 1, 2023 to Apr 30, 2024

- There are three types of behaviors being compared within the exposure vs engagement framework:
  - Google Search ***Exposure***: This uses domains occurring in the search results on a particular Google search page
  - Google Search ***Follows***: This looks at domains that users directly go to from google.com – the source of a website visit is Google search
  - Overall ***Engagement***: This simply looks at all the domains that users visit in their browser, regardless of source of the visit or other considerations

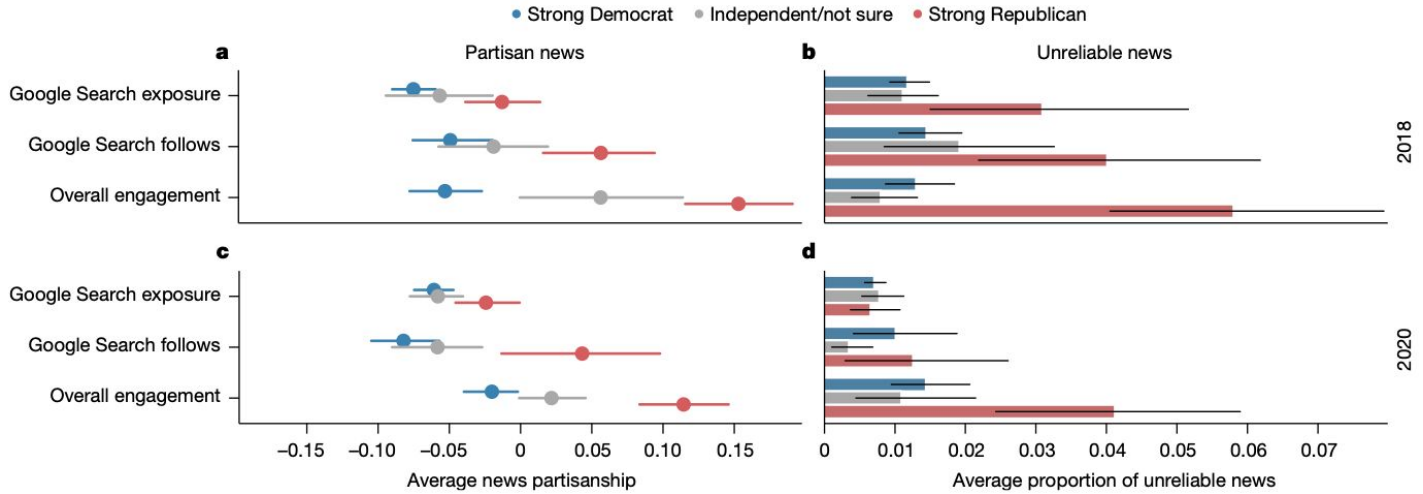
Main finding: "Strong partisans are exposed to similar rates of partisan and unreliable news, but asymmetrically follow and engage with such news"

For Robertson et al:

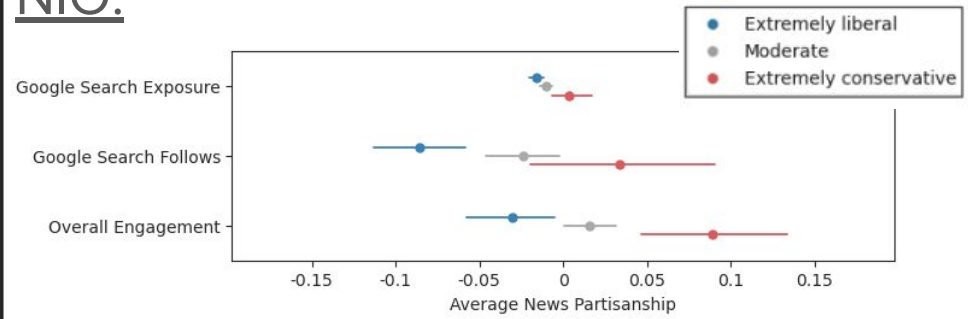


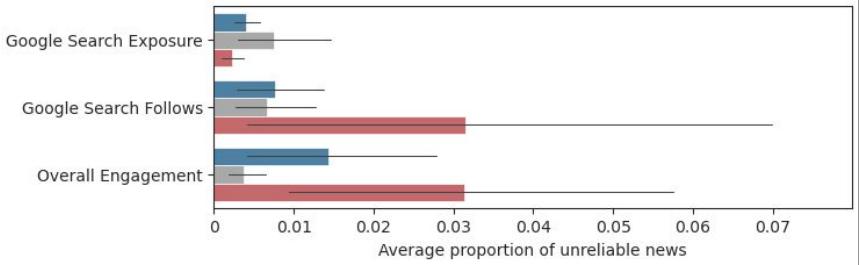
Main finding: "Strong partisans are exposed to similar rates of partisan and unreliable news, but asymmetrically follow and engage with such news"

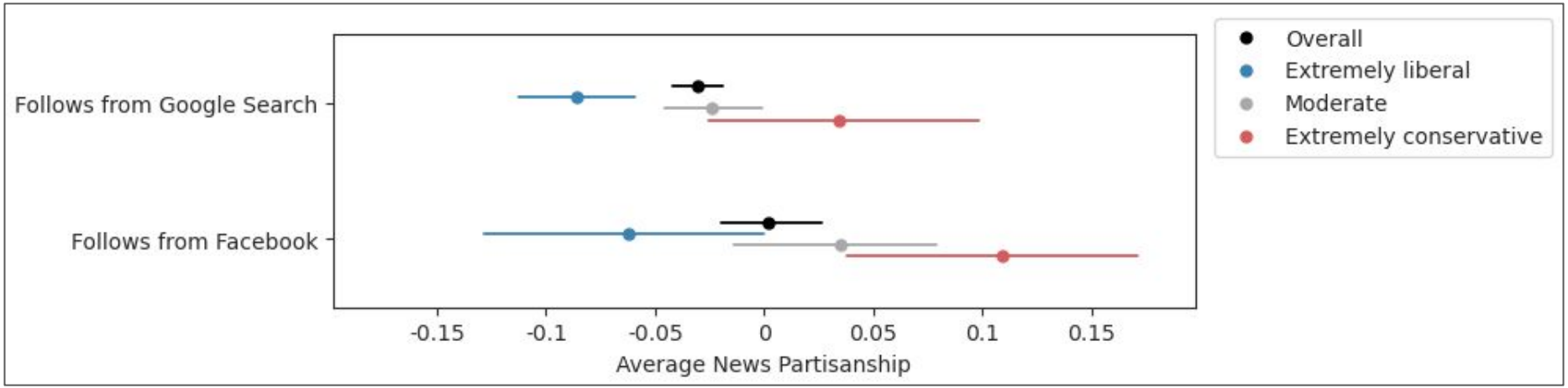
For Robertson et al:



NIO:





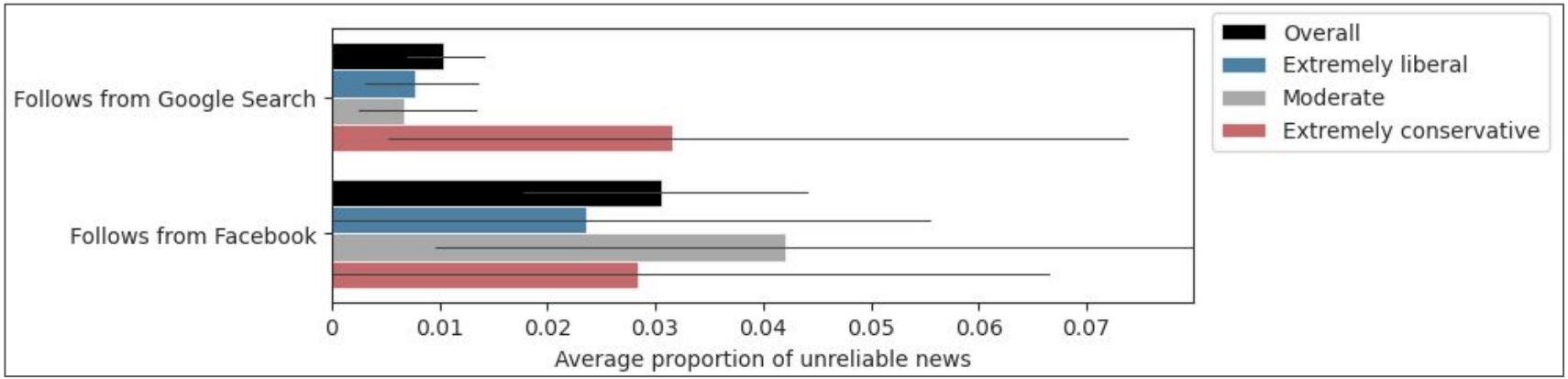


Overall, news followed from Facebook as the source is *significantly* more conservative leaning than news followed from Google search, driven by extremely conservative users and also moderate users.

For both Google and Facebook, there is a *significant* gap between average news partisanship of websites visited by strong partisans; but the gap is slightly higher for news diet originating in Facebook.

Beyond Replication:

Extending the study with Google vs Facebook *follows* data



Overall, Facebook is *significantly* more likely to direct users to unreliable news sources than Google; the proportion of unreliable news diet originating in Facebook is three times the proportion originating in Google.

While extremely conservative users are much more likely to go on to unreliable news from Google, the proportion of unreliable news originating in Facebook is more evenly distributed across extremely liberal and extremely conservative users.

Web Browsing (Link Attribution)

Name	Description	Level of Data
parser_versi	parser version	whole collection
max_link_d	maximum time permitted between link source/destination, in seconds	whole collection
uid	unique identifier for a specific user	user
blurr_versio	Blurr version	link
dest_frameI	unique identifier for destination frame	link
dest_domair	destination domain	link
dest_transiti	destination transition qualifier. We don't attribute link transitions with the "forward_back" qualifier. Furthermore, researchers may want to treat "client_redirect" or "server_redirect" transitions differently.	link
src_frameId	unique identifier for source frame	link
src_domain	source domain	link
src_transitio	source transition type	link
src_transitio	source transition qualifier	link

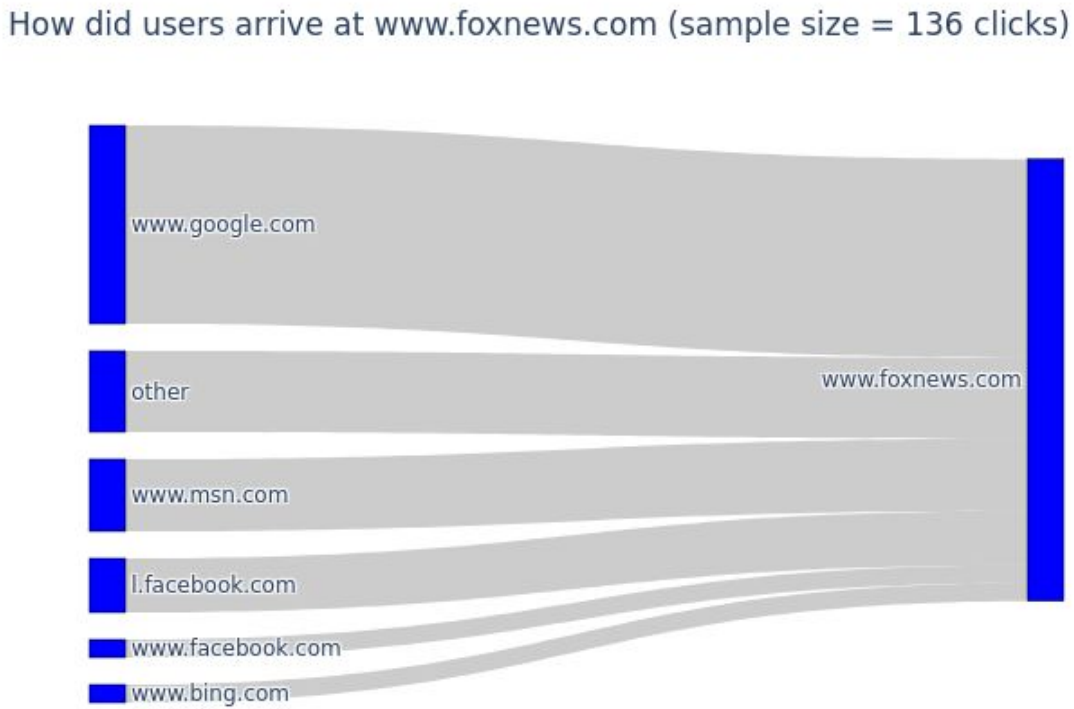
Ending point

Starting point

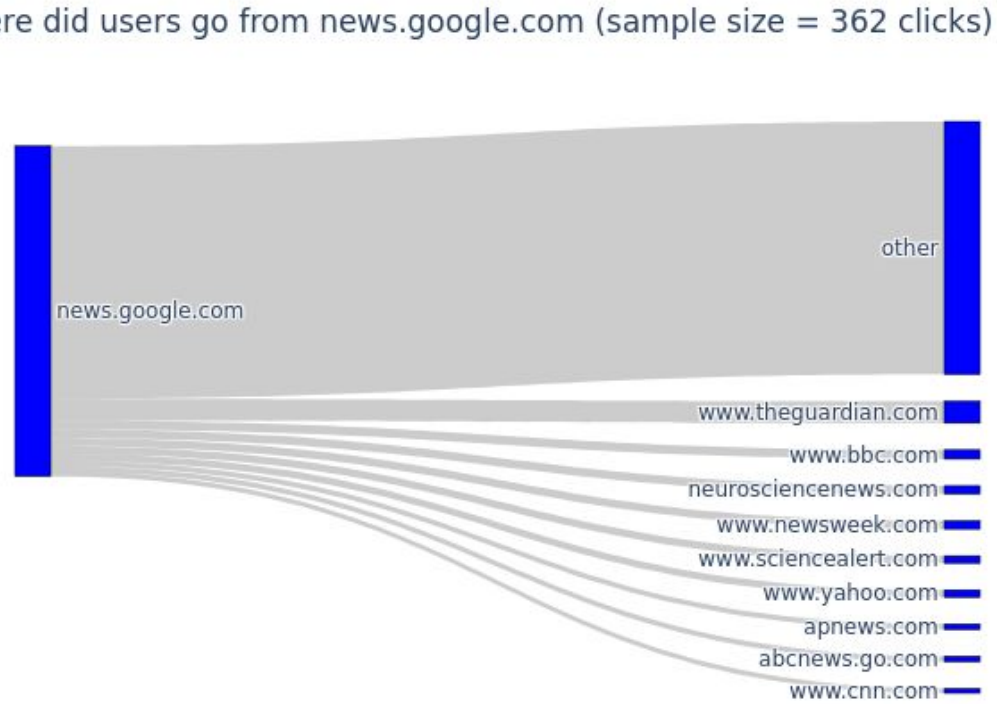
Types: link, typed, bookmark, start_page, form submission, redirect, etc.

Visits to New York Times

Visits to Fox News



Outgoing Visits from Google News



Google Search

Each row is one component on one Google SERP

Name	Description	Level of Data
scraper_version	scraper version	whole collection
parser_version	parser version	whole collection
uid	unique identifier for a specific user	user
frameId	unique identifier for a specific Google Search page	page
collectionDurationMillis	milliseconds elapsed during scraping, may be useful for filtering errors.	page
qry	query	page
tbm	tbm parameter in URL, identifies vertical search (e.g. news is `nws`, shopping is `shop`, local is `lcl`)	page
infinite_scroll	whether this result was generated after scrolling; we do not currently parse these results, so all columns are `None`	page
type	component type (e.g. `knowledge`, `top_stories`)	component
cmpt_rank	component vertical rank (main results column first, then right-hand side panel)	component
sub_rank	component horizontal rank (e.g. within a `top_stories` carousel)	component
serp_rank	overall component rank, reading from left to right, top to bottom	component
url	component URL	component
title	component title	component
cite	component citation	component
details	details included for some component types	component
text	component text snippet	component
sub_type	sub-type for some component types (e.g. is a `knowledge` component a `featured_snippet` or a `calculator`)	component
rhs_column	whether this result is part of the right-hand side column	component
timestamp	timestamp included in some component types (e.g. `twitter_results`, `news_quotes`)	component
error	parsing error stack trace	component