Lecture 13: Experiments; Human Subjects

CMSC 25910

Winter 2026
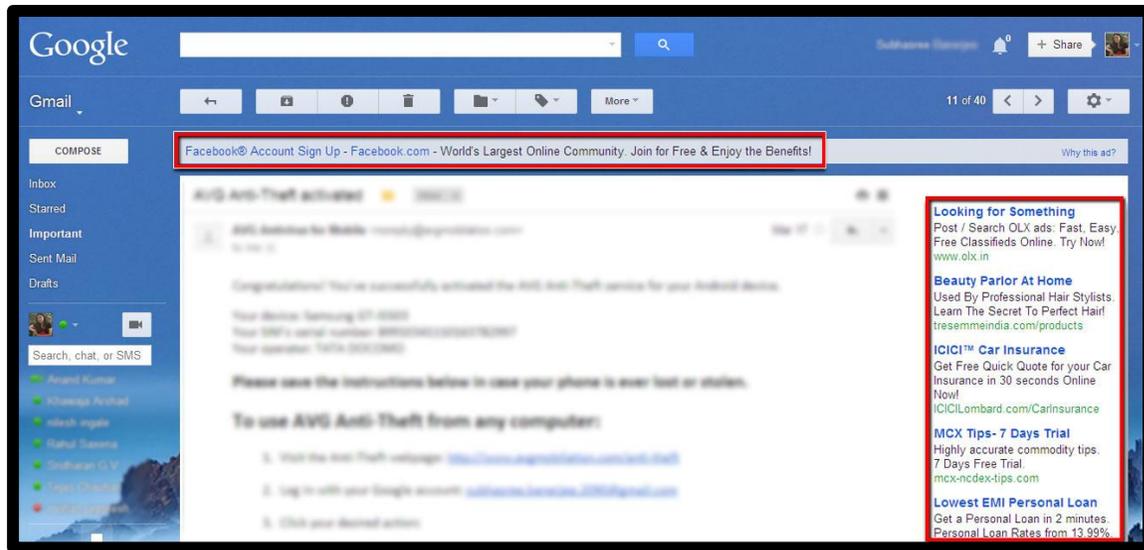
The University of Chicago

THE UNIVERSITY OF CHICAGO

Why We Run Experiments

A/B Testing & Data-Based Experiments

Google's commitment to data-driven decisions is well reported, and the company has been ridiculed for the "50 shades of blue" episode, when then Google executive Marissa Meyer led a project testing the impact of using different coloured links in ads.

But a new insight proves that the company significantly benefitted from the experiment, to the tune of $200m.

The figure comes from Google UK's managing director Dan Cobley, speaking on Tuesday at an event organised by law firm DLA Piper, who positioned the company's approach to data against the traditional route of the "highest paid person's opinion".

"About six or seven years ago, Google launched ads on Gmail," Cobley explained. "In our search we have ads on the side, little blue links that go to other websites: we had the same thing on gmail. But we recognised that the shades of blue in those two different products were slightly different when they linked to ads.

"In the world of the hippo, you ask the chief designer or the marketing director to pick a blue and that's the solution. In the world of data you can run experiments to find the right answer.

"We ran '1%' experiments, showing 1% of users one blue, and another experiment showing 1% another blue. And actually, to make sure we covered all our bases, we ran forty other experiments showing all the shades of blue you could possibly imagine.

"And we saw which shades of blue people liked the most, demonstrated by how much they clicked on them. As a result we learned that a slightly purpler shade of blue was more conducive to clicking than a slightly greener shade of blue, and gee whizz, we made a decision.

"But the implications of that for us, given the scale of our business, was that we made an extra $200m a year in ad revenue."

The form of testing Google undertook is known as A/B testing (offering users two different versions of a site and picking the most effective one); this particular battle was widely seen as a turning point for the company, the moment it sided with engineers against designers. In 2009, Doug Bowman, then the company's top designer, cited it as part of the reason for his departure.

Alex Hern. Why Google has 200m reasons to put engineers over designers. The Guardian.
https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers

A/B Testing & Data-Based Experiments

- Randomly / systematically vary a quantity of interest

- Measure some sort of quantifiable outcome

- Make decisions from data
  - *Is this always the right way?*

Google's commitment to data-driven decisions is well reported, and the company has been ridiculed for the "50 shades of blue" episode, when then Google executive Marissa Meyer led a project testing the impact of using different coloured links in ads.

But a new insight proves that the company significantly benefitted from the experiment, to the tune of $200m.

The figure comes from Google UK's managing director Dan Cobley, speaking on Tuesday at an event organised by law firm DLA Piper, who positioned the company's approach to data against the traditional route of the "highest paid person's opinion".

"About six or seven years ago, Google launched ads on Gmail," Cobley explained. "In our search we have ads on the side, little blue links that go to other websites: we had the same thing on gmail. But we recognised that the shades of blue in those two different products were slightly different when they linked to ads.

"In the world of the hippo, you ask the chief designer or the marketing director to pick a blue and that's the solution. In the world of data you can run experiments to find the right answer.

"We ran '1%' experiments, showing 1% of users one blue, and another experiment showing 1% another blue. And actually, to make sure we covered all our bases, we ran forty other experiments showing all the shades of blue you could possibly imagine.

"And we saw which shades of blue people liked the most, demonstrated by how much they clicked on them. As a result we learned that a slightly purpler shade of blue was more conducive to clicking than a slightly greener shade of blue, and gee whizz, we made a decision.

"But the implications of that for us, given the scale of our business, was that we made an extra $200m a year in ad revenue."

The form of testing Google undertook is known as A/B testing (offering users two different versions of a site and picking the most effective one); this particular battle was widely seen as a turning point for the company, the moment it sided with engineers against designers. In 2009, Doug Bowman, then the company's top designer, cited it as part of the reason for his departure.

Alex Hern. Why Google has 200m reasons to put engineers over designers. The Guardian.
https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers

The (Non-)Protection of Human Subjects

A Historical Perspective

The Monster Study (1939)

- University of Iowa, Wendell Johnson and Mary Tudor
  - RQ: Impact of affirmation/criticism in speech therapy
- Participants: 22 orphan children
  - One group received positive speech therapy and praise for fluency
  - Another group received criticism for every imperfection
- Children in the 2nd group developed permanent speech issues

The Milgram Study (1961)

- Yale University, Stanley Milgram
  - RQ: Obedience of authority figures
- Participants shocked another "participant" (actually an actor who was a confederate of the experimenters) for wrong answers
- Most subjects expressed a desire to stop, but continued when told they would not be held responsible

Stanford Prison Experiment (1971)

- Stanford psychology professor Philip Zimbardo
  - RQ: Is brutality a personality trait of guards or situational?

- Participants: 24 men recruited for 2-week experiment
  - 12 took role of *prisoners,* assigned numbers and uniforms
  - 12 took role of *guards* with wooden batons and uniforms

- Prisoners were arrested at home, stripped naked at the "prison"

- Rebellion, degradation, breakdowns occurred

- Physical and psychological trauma
  - Experiment terminated after 6 days

Tuskegee Syphilis Experiment (1972)

- US Public Health Service and Tuskegee University *"Tuskegee Study of Untreated Syphilis in the Negro Male"* 1932-1972

- Participants: 600 impoverished, African-American sharecroppers
  - 399 with syphilis, 201 without (control group)

- Not told they had syphilis

- Not treated with penicillin
  - Notably, syphilis was entirely treatable by the end of the experiment

Principles For
Protecting Human Subjects

Belmont Report

- *"Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research"*
- September 30, 1978
- Three key principles (following slides)

Principle 1: Respect for Persons

- Protect the autonomy of persons

- Informed consent

- No deception* (will revisit later in this lecture)

Principle 2: Beneficence

- "Do no harm"
- Maximize benefits relative to risks

Principle 3: Justice

- Do not exploit participants
- Fairly distribute costs/benefits to prospective participants

Menlo Report

- The Menlo Report (2012) focused on the intersection between human-subjects experiments and cybersecurity research
- It added a fourth principle: respect for the law and public interest
- https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf

Ethical Principles in Security Research

- Responsible disclosure of vulnerabilities
- Potential requests for permission to test (contested!)
- Respect for direct human subjects in research/experiments
- Respect for humans indirectly affected
- Careful analysis about whether to experiment on real systems
- Evaluation of the impacts on running systems/networks
- Possible consideration of terms of service (contested!)
- Evaluation of the downstream impacts of the research on all possible stakeholders

The Continued (Non-)Protection of Human Subjects

Recent CS Perspectives

Social Contagion Precursor (2012)

The Spread of Emotion via Facebook

Adam D. I. Kramer

Facebook, Inc.

1601 Willow Rd, Menlo Park, CA

akramer@fb.com

ABSTRACT

In this paper we study large-scale emotional contagion through an examination of Facebook status updates. After a user makes a status update with emotional content, their friends are significantly more likely to make a valence-consistent post. This effect is significant even three days later, and even after controlling for prior emotion expressions by both users and their friends. This indicates not only that emotional contagion is possible via text-only communication and that emotions flow through social networks, but also that emotion spreads via indirect communications media.

Author Keywords

Computer-mediated communication; emotional contagion; emotion; Facebook; social networks

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

increasingly supported via CMC. Facebook, a very popular SNS claiming over 800M users as of publication, is a common means by which people interact with each other both directly (such as via messages, chatting, wall posts, comments), and indirectly (such as via status updates).

In this paper, we examine an important implication of emotional contagion theory: whether and how emotions "spread" via SNSs. The examination of how emotions spread through a textual medium such as an internet SNS is very interesting: If such a process can be shown, this provides evidence for basic psychological processes re-exhibiting themselves in new media, and also contributes to a growing body of evidence suggesting that interaction with "friends" on SNSs mirrors the manner by which people interact with others in offline life [4,10].

Though the current studies are correlational, they address several confounds in prior results by making use of the undirected nature of Facebook "status updates:" Prior

Social Contagion Precursor (2012)

The Spread of Emotion via Facebook

Adam D. I. Kramer

Facebook, Inc.

1601 Willow Rd, Menlo Park, CA

akramer@fb.com

ABSTRACT

In this paper we study large-scale emotional contagion through an examination of Facebook status updates. After a user makes a status update with emotional content, their friends are significantly more likely to make a valence-consistent post. This effect is significant even three days later, and even after controlling for prior emotion expressions by both users and their friends. This indicates not only that emotional contagion is possible via text-only communication and that emotions flow through social networks, but also that emotion spreads via indirect communications media.

Author Keywords

Computer-mediated communication; emotional contagion; emotion; Facebook; social networks

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

increasingly supported via CMC. Facebook, a very popular SNS claiming over 800M users as of publication, is a common means by which people interact with each other both directly (such as via messages, chatting, wall posts, comments), and indirectly (such as via status updates).

In this paper, we examine an important implication of emotional contagion theory: whether and how emotions "spread" via SNSs. The examination of how emotions spread through a textual medium such as an internet SNS is very interesting: If such a process can be shown, this provides evidence for basic psychological processes re-exhibiting themselves in new media, and also contributes to a growing body of evidence suggesting that interaction with "friends" on SNSs mirrors the manner by which people interact with others in offline life [4,10].

Though the current studies are correlational, they address several confounds in prior results by making use of the undirected nature of Facebook "status updates:" Prior

Facebook Social Contagion (2014)

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are later seen by their friends via Facebook's "News Feed" product (8). Because people's friends frequently produce much more content than one person can view, the News Feed filters posts, stories, and activities undertaken by friends. News Feed is the primary manner by which people see content that friends share. Which content is shown or omitted in the News Feed is de-

Facebook Social Contagion (2014)

- "We show, via a massive (N = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues."
https://www.pnas.org/content/111/24/8788

Facebook Social Contagion (2014)

CORRECTION

Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks

VIEW RELATED CONTENT +

PDF

PSYCHOLOGICAL AND COGNITIVE SCIENCES PNAS is publishing an Editorial Expression of Concern regarding the following article: "Experimental evidence of massive-scale emotional contagion through social networks," by Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, which appeared in issue 24, June 17, 2014, of *Proc Natl Acad Sci USA* (111:8788–8790; first published June 2, 2014; 10.1073/pnas.1320040111). This paper represents an important

Linux Hypocrite Commits (2021)

On the Feasibility of Stealthily Introducing Vulnerabilities in Open-Source Software via Hypocrite Commits

Qiushi Wu and Kangjie Lu
University of Minnesota
{wu000273, kjlu}@umn.edu

Abstract—**Open source software (OSS) has thrived since the forming of Open Source Initiative in 1998. A prominent example is the Linux kernel, which has been used by numerous major software vendors and empowering billions of devices. The higher availability and lower costs of OSS boost its adoption, while its openness and flexibility enable quicker innovation. More importantly, the OSS development approach is believed to produce more reliable and higher-quality software since it typically has thousands of independent programmers testing and fixing bugs of the software collaboratively.**

In this paper, we instead investigate the insecurity of OSS from a critical perspective—the feasibility of stealthily introducing vulnerabilities in OSS via hypocrite commits (i.e., seemingly beneficial commits that in fact introduce other critical issues).

Its openness also encourages contributors; OSS typically has thousands of independent programmers testing and fixing bugs of the software. Such an open and collaborative development not only allows higher flexibility, transparency, and quicker evolution, but is also believed to provide higher reliability and security [21].

A prominent example of OSS is the Linux kernel, which is one of the largest open-source projects—more than 28 million lines of code used by billions of devices. The Linux kernel involves more than 22K contributors. Any person or company can contribute to its development, e.g., submitting a patch through git commits. To make a change of the Linux kernel,

Linux Hypocrite Commits (2021)

ZDNET — tomorrow belongs to those who embrace it today

/ business

Home / Business / Enterprise Software

Greg Kroah-Hartman bans University of Minnesota from Linux development for deliberately buggy patches
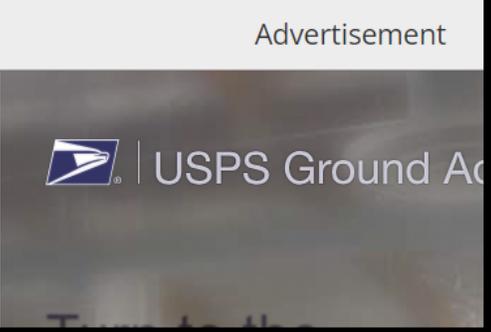
tom's HARDWARE

US Edition

Software > Operating Systems > Linux

University Responds to Ban On Linux Contributions

News By Nathaniel Mott published April 22, 2021

Advertisement

USPS Ground A

Princeton GDPR/CCPA Study (2021)

- Research question: How have online services implemented GDPR and CCPA data subject access rights?
  - "Our goals are to accurately describe how websites have operationalized these new user rights, whether websites are extending these rights to non-EU citizens and non-California residents, and whether websites are effectively authenticating users when they exercise these rights."
- Study design: Send automated emails from (fictional) people to a random sample of websites inquiring about data access rights

Princeton GDPR/CCPA Study (2021)

To Whom It May Concern:

My name is [REDACTED], and I am a resident of Sacramento, California. I have a few questions about your process for responding to General Data Protection Regulation (GDPR) data access requests:

Would you process a GDPR data access request from me even though I am not a resident of the European Union?

Do you process GDPR data access requests via email, a website, or telephone? If via a website, what is the URL I should go to?

What personal information do I have to submit for you to verify and process a GDPR data access request?

What information do you provide in response to a GDPR data access request?

To be clear, I am not submitting a data access request at this time. My questions are about your process for when I do submit a request.

Thank you in advance for your answers to these questions. If there is a better contact for processing GDPR requests regarding [DOMAIN], I kindly ask that you forward my request to them.

I look forward to your reply without undue delay and at most within one month of this email, as required by Article 12 of GDPR.

Sincerely,[REDACTED]

https://www.reddit.com/r/gdpr/comments/reqlxs/questions_about_gdprccpa_data_access_process_scam/

Princeton GDPR/CCPA Study (2021)

coywolf.news/webdev/princeton-radboud-researchers-halt-gdpr-ccpa-study-over-ethics-concerns-and-industry-blowback/

/ News / Web Dev

Princeton privacy study halts GDPR/CCPA research over ethics concerns and industry blowback

Computer science researchers at Princeton University and Radboud University conducted an academic study that sent automated email messages to websites with GDPR and CCPA privacy policy requests from fake personas. Email operators, web admins, and privacy professionals interpreted the messages as security risks and legal threats, prompting the researchers to suspend the study and delete all communication.

By Jon Henshaw
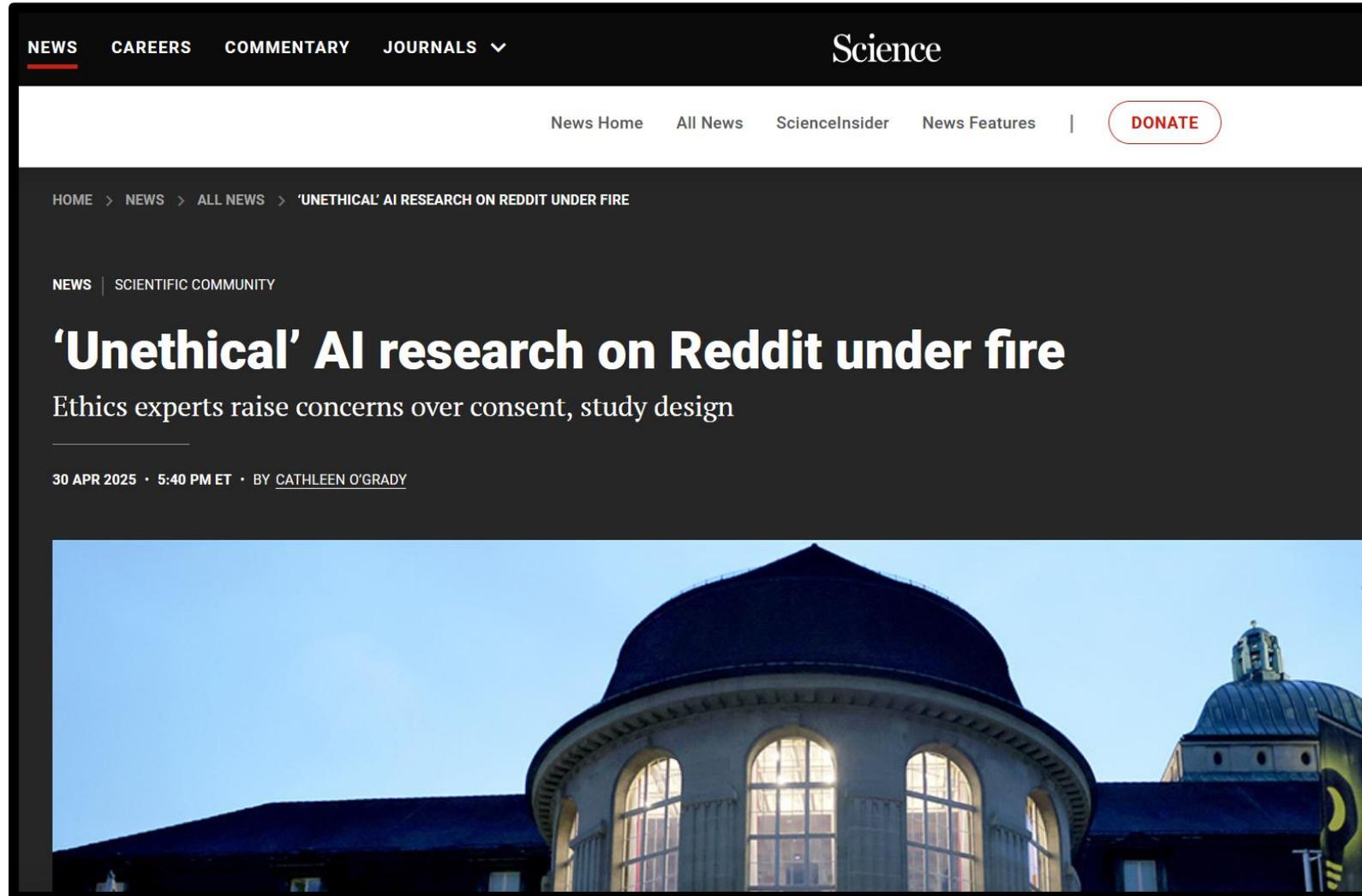12/27/2021 (Last updated on 11/3/2022)

BOING BOING / SEARCH / STORE / MENU

Princeton study tricked small websites into thinking they could be sued by a Russian organization

MARK FRAUENFELDER / 9:23 AM THU DEC 23, 2021
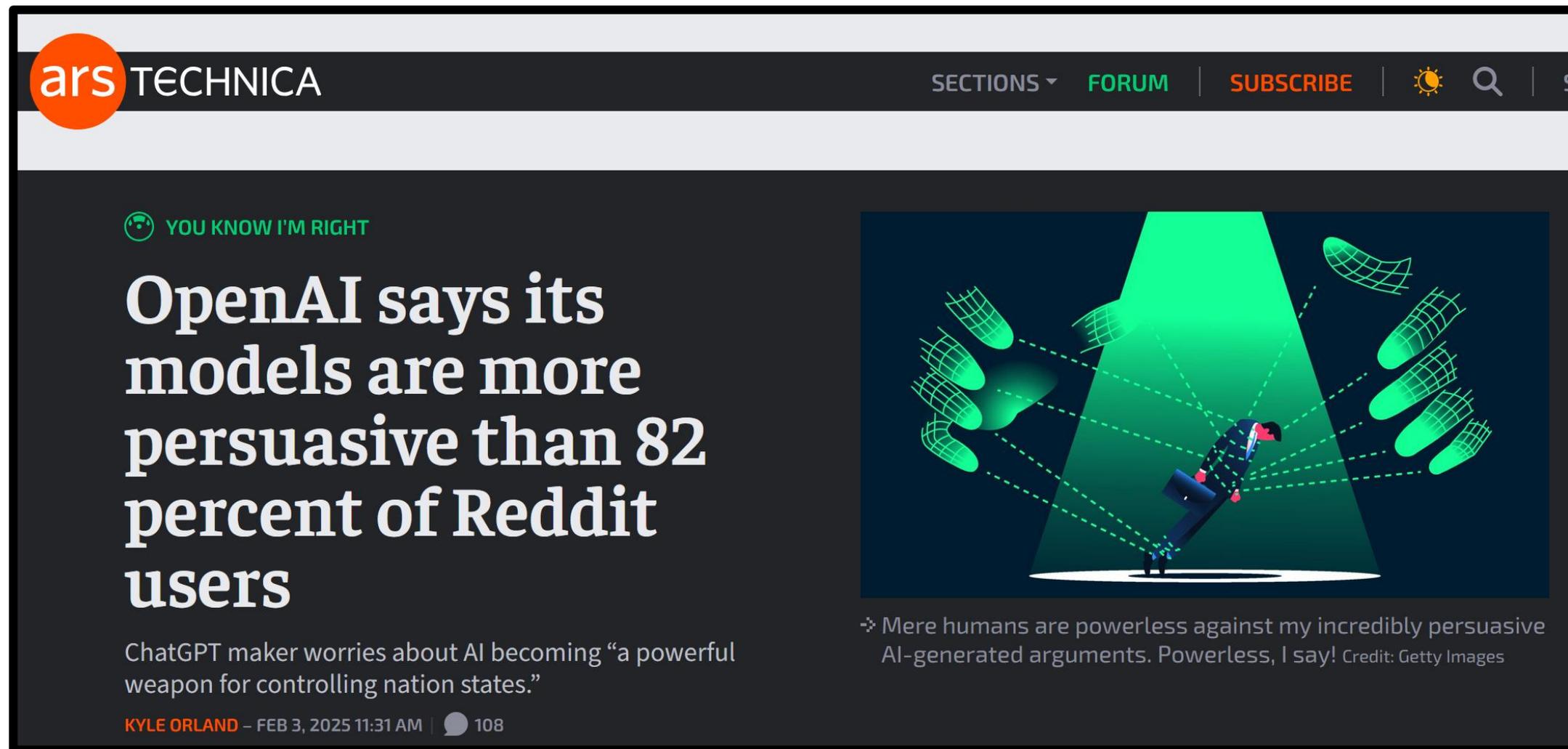
University of Zurich Reddit Study (2025)

- Research question: Can AI-generated posts on Reddit's "Change My View" subreddit change human users' views?

- Study design: "They used a range of LLMs to generate responses to discussions on r/changemyview, in some cases tailoring the arguments to individual participants by inferring demographic features such as their gender, ethnicity, and political orientation from their posting histories. The comments from the LLMs also purported to be from a range of human identities, including a male rape survivor, a trauma counselor, and a Black person who disagreed with the Black Lives Matter movement. In all, the researchers created 34 accounts and posted more than 1500 times on the platform."

Quote from https://www.science.org/content/article/unethical-ai-research-reddit-under-fire

University of Zurich Reddit Study (2025)

Quote from https://www.science.org/content/article/unethical-ai-research-reddit-under-fire

Compare to: OpenAI (2025)

ars TECHNICA

SECTIONS FORUM | SUBSCRIBE | ☀ 🔍 | S

YOU KNOW I'M RIGHT

OpenAI says its models are more persuasive than 82 percent of Reddit users

ChatGPT maker worries about AI becoming "a powerful weapon for controlling nation states."

KYLE ORLAND – FEB 3, 2025 11:31 AM | 💬 108

✧ Mere humans are powerless against my incredibly persuasive AI-generated arguments. Powerless, I say! Credit: Getty Images

The Nuances of Participants, Consent, and Deception

Participants

- Recruit people to do something remotely (e.g., online)
- Recruit people to come to your lab
- Recruit people to let you into their "context"
- Observe people (if possible, get consent! If not possible, consider necessity of design)

Participants

- What recruitment mechanisms?
  - Craigslist, flyers, participant pools, representative sample, standing on street
- How do you compensate them?
  - Ethics of paying $0.00 vs. $10.00 vs. $100,000
- How do you get informed consent?
- What happens to their data?
- Prior knowledge / "what" are they?

Ethical Questions to Ask Yourself

- How do we protect participants?
  - What risks do we introduce?
- Is there a less invasive method that would give equivalent insight?
- How do we make sure participation is voluntary throughout the experiment?

Institutional Review Board (IRB)

- Is it research? Are there human subjects?
- IRB is one arbiter of ethics; experimenters themselves are another crucial arbiter
- Full review vs. expedited vs. exempt
- Fill out and submit protocol
  - Include all study materials (e.g., surveys)
  - Include recruitment text and/or poster
  - Leave plenty of time

What to submit to an IRB

- Full consent form (use UChicago model)
- All scripts, survey questions, instructions
- Recruitment plan
- Recruitment materials
  - Don't emphasize compensation
- Information about how data will be handled
  - Password protection, encryption, etc.
  - Meetings to discuss

Informed Consent Templates

https://sbsirb.uchicago.edu/templates/

THE UNIVERSITY OF CHICAGO

Version: [e.g.,1.0]

Consent Form for Research Participation

Study Number: e.g., IRB18-XXXX
Study Title: List the title or short title as provided in the application
Researcher(s): List at least the PI
Sponsor: (if applicable, or remove)
Collaborating Institution(s): (if applicable, or remove)

This is a consent form for research participation. It contains important information about this study and what to expect if you decide to participate. Your participation is voluntary.

Purpose: Explain why the research is being done

Procedures and Time Required: For example, "You will be asked to participate in two 30-minute interviews over the phone. With your permission, the interviews will be audio-recorded."

Financial Information: Please explain the amount and terms of any payments or reimbursements. If payments will be prorated if a subject withdraws from the study, explain. If including a raffle or lottery, be sure to include the required language listed in the supplemental consent template language document. If this section is not applicable, remove or state "Participation in this study will involve no cost to you. You will not be paid for participating in this study."

Informed Consent Templates

https://sbsirb.uchicago.edu/templates/

Risks and Benefits: As applicable. If no direct risks, indicate, "Your participation in this study does not involve any risks to you beyond those of everyday life." If no direct benefits to individuals, you should indicate if there are potential benefits to others; e.g., "Taking part in this research study may not benefit you personally, but we may learn new things that could help others."

If applicable, explain any alternatives to participation, especially if research involves a program, treatment, or therapy.

Confidentiality: Describe how data, recordings, identifiers (if any), etc. will be used, shared, and protected during the research, as well as their retention, use, or disposition following the conclusion of the research.

- As applicable, address how partially collected data will be handled in case of a participant withdraws: e.g., "If you decide to withdraw from this study, the researchers will ask you if the information already collected from you can be used," "If you decide to withdraw, data collected up until the point of withdrawal may still be included in analysis," or "If you decide to withdraw from this study, any data already collected will be destroyed."
- Indicate whether identifiable data may be shared for future research, e.g., "Identifiable information will be handled as described in the 'optional elements section' below," or "Identifiable data will never be shared outside the research team."
- Address whether de-identified data may be shared for future research, e.g., "De-identified information from this study may be used for future research studies or shared with other researchers for future research without your additional informed consent," or,

Informed Consent Templates

https://sbsirb.uchicago.edu/templates/

"The information collected as part of this research will not be used or shared for future research studies, even if all identifiers are removed."
- If applicable, add mandated reporter language.
- Note: if the study is NIH-funded, please be sure to include the CoC language listed in the supplemental consent template language document if required.

Contacts & Questions:
If you have questions or concerns about the study, you can contact the researchers at [add your contact information, including name, telephone number, and email address].

If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research team, you can contact the University of Chicago Social & Behavioral Sciences Institutional Review Board (IRB): phone (773) 702-2915, email sbs-irb@uchicago.edu.

Consent:
Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled. You will be provided a copy of this form. By signing below, you agree to participate in the research.

_____
Participant's Signature

_____   _____
Participant's Name (printed)   **Date**

Deception

- Do we mind if participants know precisely what is being studied?
  - Sometimes, it's crucial that we observe their organic responses in context
- What "deception" or "distraction" task can we introduce?
- How do we **debrief** people at the end?

Designing an Experiment

Defining the purpose and goals

- What are you hoping to learn?
  - That is, what are your research questions?
  - **Precisely stated research questions are crucial**
- What are your explicit hypotheses?
- What are your metrics?
  - What data might be directly or indirectly helpful?
- What, if anything, are you comparing to?
  - Control condition or baseline

Research questions (RQs)

- Succinct, **precisely stated**
  - Generally a **falsifiable statement** or **specific question**
  - Usually, but not always, encodes some sort of hypothesis
- Goals of the research can be broad, whereas RQs are usually more narrow
- Let your RQs guide the design of your experiment

Broad types of studies

- Formative (initial) vs. summative (validation)
- **Descriptive study**: describe a phenomenon
- **Relational study**: correlation between variables
- **Experimental study**: causation

- We can do studies with **humans**
  - How humans use a system
  - System-relevant characteristics of humans
- We can also study **systems themselves**

STAND BACK

I'M GOING TO TRY
SCIENCE

Quantitative vs. Qualitative

- **Quantitative**: numbers
  - Timing how long we awkwardly wait for you all to answer a question
  - Ratings of the course staff's awesomeness on a numerical scale
  - How long it took a computational process to complete
- **Qualitative**: non-numerical data
  - Free-text thoughts, opinions, understanding, types of errors

Roles for data

- **Independent variables**: explanatory variables
  - Which variant/condition/treatment was assigned
  - **Covariates**: characteristics of participants (demographics, experiences, or other aspects) that could explain differences

- **Dependent variable(s)**: your main metric(s) of interest
  - The primary thing you're measuring that you expect to change based on the variant/condition/treatment

Point(s) of comparison

- Control condition / baseline:
    - May be a **placebo** (no actual intervention)…
    - …or it may be a state-of-the-art system
- You often create variants (**conditions, treatments**)
    - Think about your research questions and how comparing pairs or groups of conditions lets you answer your research questions
    - A common rookie mistake is not having sensibly matched sets of conditions, introducing **confounds** (other factors that might cause any differences observed)
- When studying a system, you might need "typical" (or intentionally atypical) workloads or traces

Study designs

- **Within-subjects**
  - Every participant tests everything
  - Crucial to randomize order! (learning effect)
  - Fewer participants needed, but longer study
- **Between-subjects**
  - Each participant tests 1 version of the system
  - You compare these groups
  - Groups should be similar (verify!)
  - Still randomize!

Conclusions From Studies

Validity

- To what degree are we confident that X causes Y (**internally valid**)?

- To what degree can we generalize about our results (**externally valid**)?
    - What biases does our sample introduce?

- Is this study **ecologically valid**?
    - Does it mirror real-life conditions and context?

- Balancing all of these is hard!

What we conclude from studies

- It's very rare that we conclude something like "for all humans there is an X% effect of Y" or "Z% of people care about ethics"
  - Be clear about what population you have sampled
- We often use proxies in measurement

What we conclude long-term

- **Repeatability**: findings consistent with same researchers and same infrastructure

- **Reproducibility**: findings consistent with different researchers and different (comparable) infrastructure

- Sadly, few studies are replicated
  - Bias against successful replication in peer review
  - (Also) bias against publishing negative results

Some potential confounds

- Measurement accuracy / resolution
- Differences caused by different experimental platforms
- Order of recruiting matters
  - Round-robin (123123123, etc.), Latin squares
- Time of day for recruiting matters
- Failing to account for study dropout or non-participation (very subtle!)
- Changing multiple aspects across conditions that are compared

Some potential confounds

- Learning effect
  - Randomize order of tasks
  - Consider learning effect as a covariate
- Different instructions for different participants
- Biases of recruitment / representativeness
- Self-report biases
  - Don't ask people to rate expertise

Some potential confounds

- Different demographics in conditions
- **Placebo effect**
  - Why you need a control condition
- **Hawthorne effect** (changing behavior in response to being observed)
- Habituation / novelty
  - People pay more attention to new things
- Participants try to please experimenter
  - I like yours better!
  - Minimize knowledge of what's being tested

Process

An example study

- (Bad) research question: "Is UChicago the place where fun comes to die?"

- Recruiting participants: what can go wrong?

- Independent variable: assign a university?

- Dependent variable: some proxy for fun
  - Hours not studying?
  - Hours not in the Reg?
  - Agreement with statement "We are having fun"

Another example study

- What if you have a computationally expensive ML-based intrusion detection system you have created to detect network-based attacks

- What are your research questions?

- What are your variables of interest?

Describing your methods

- Be clear and honest about what you did
  - Be honest, earnest, and upfront about limitations
- Give enough detail for someone to replicate
  - Study materials as appendix if possible
  - Correctly report stats (e.g., APA guidelines)
- Release code if possible
- Release data if possible
  - Requires approval from IRB **and** participants