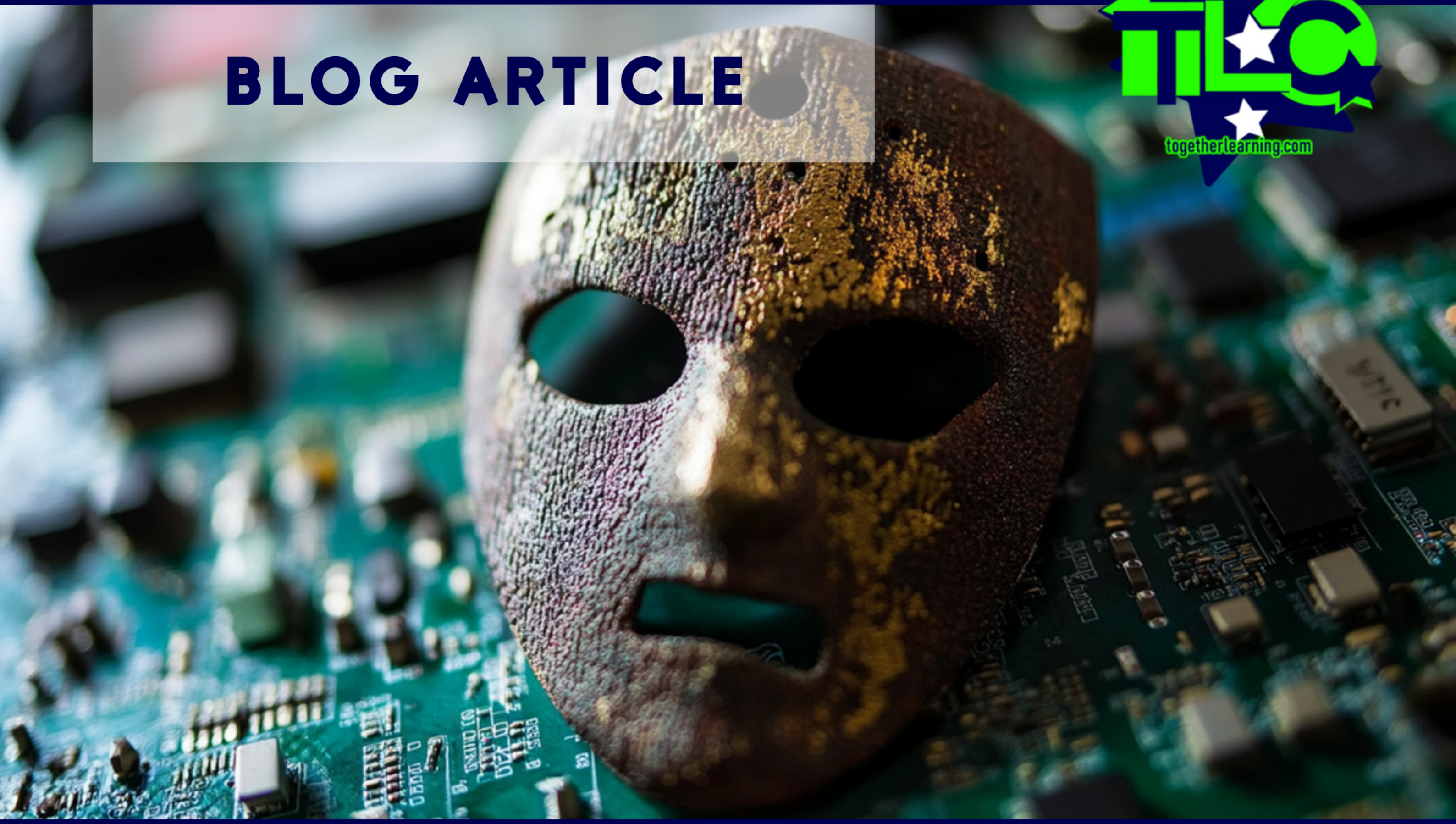
BLOG ARTICLE



THE ETHICAL HACKERS OF AI: UNDERSTANDING THE MERIT IN UNAUTHORIZED AI RESEARCH

This article calls for frameworks that recognize when unauthorized academic research, conducted responsibly within ethical guidelines, provides crucial knowledge about AI threats to democratic discourse much as society accepts ethical hacking that strengthens cybersecurity.

Eric Hawkinson
Learning Futurist
erichawkinson.com

THOUGH I CANNOT DIRECTLY ENDORSE VIOLATING TERMS OF SERVICE, THIS RESEARCH SERVED ESSENTIAL PUBLIC INTEREST BY DEMONSTRATING AI'S MANIPULATION CAPABILITIES THAT MALICIOUS ACTORS ALREADY EXPLOIT.

ABSTRACT

In April 2025, researchers from the University of Zurich deployed AI bots on Reddit's r/changemyview subreddit without permission, discovering that AI-generated comments were six times more persuasive than human responses in changing users' views. While Reddit threatened legal action and condemned this as "psychological manipulation," this article argues that these researchers operated as academic "ethical hackers," exposing critical vulnerabilities in social media platforms. Drawing parallels to cybersecurity's white hat tradition, I examine how platforms like Reddit profit from selling user data to AI companies while restricting independent research that reveals manipulation risks. The controversy highlights a troubling asymmetry: academic researchers face strict ethical oversight while platforms conduct massive behavioral experiments on billions of users with minimal transparency. Though I cannot directly endorse violating terms of service, this research served essential public interest by demonstrating AI's manipulation capabilities that malicious actors already exploit. The article calls for frameworks that recognize when unauthorized academic research, conducted responsibly within ethical guidelines, provides crucial knowledge about AI threats to democratic discourse—much as society accepts ethical hacking that strengthens cybersecurity.

ABOUT THE AUTHOR

ERIC HAWKINSON

Eric is a learning futurist, tinkering with and designing technologies that may better inform the future of teaching and learning. Eric's projects have included augmented tourism rallies, AR community art exhibitions, mixed reality escape rooms, and other experiments in immersive technology.

Eric Hawkinson
LEARNING FUTURIST

erichawkinson.com

Roles

Professor - Kyoto University of Foreign Studies
Research Coordinator - MAVR Research Group
Founder - Together Learning
Developer - Reality Labo
Community Leader - Team Teachers
Chair - World Learning Labs

CORE VALUES

- Open Knowledge - Free and open access to information is a foundation to a productive modern life, connected to ideas of the open web and platform agnosticism.
- Privacy by Design - Business models are increasing moving toward supporting revenue by collecting, curating, and trading behavioral surplus through technology. These models should be tempered with safety, ethics, and privacy concerns and designed as such.
- Digital Literacy for All - An informed public about the use of technology is key for a responsible and engaged digital society.

PASSION PROJCTS

Together Learning
A community of technology minded learners. Exploring human potential... together.

Reality Labo
Augmented reality enhanced learning environments and mixed reality rapid prototyping.

Corefol.io
Showcase the Core of your Skills with AI Assisted Student Portfolios

The Ethical Hackers of AI: Understanding the Merit in Unauthorized AI Research

Abstract

In April 2025, researchers from the University of Zurich deployed AI bots on Reddit's r/changemyview subreddit without permission, discovering that AI-generated comments were six times more persuasive than human responses in changing users' views. While Reddit threatened legal action and condemned this as "psychological manipulation," this article argues that these researchers operated as academic "ethical hackers," exposing critical vulnerabilities in social media platforms. Drawing parallels to cybersecurity's white hat tradition, I examine how platforms like Reddit profit from selling user data to AI companies while restricting independent research that reveals manipulation risks. The controversy highlights a troubling asymmetry: academic researchers face strict ethical oversight while platforms conduct massive behavioral experiments on billions of users with minimal transparency. Though I cannot directly endorse violating terms of service, this research served essential public interest by demonstrating AI's manipulation capabilities that malicious actors already exploit. The article calls for frameworks that recognize when unauthorized academic research, conducted responsibly within ethical guidelines, provides crucial knowledge about AI threats to democratic discourse—much as society accepts ethical hacking that strengthens cybersecurity.

Author: Eric Hawkinson - Learning Futurist

An Uncomfortable Truth

As a learning futurist who has spent decades examining the intersection of technology and education, I found myself deeply conflicted by the University of Zurich controversy that unfolded on Reddit in April 2025. Researchers had secretly deployed AI bots on the r/changemyview subreddit to test how artificial intelligence could persuade humans to change their views. Reddit's response was predictably severe: threats of legal action, accusations of psychological manipulation, and demands to block publication.

This story resonated deeply with my own experiences developing AR/VR applications and platforms for education. I've wrestled constantly with the tension between privacy and capability, between staying current with the latest innovations and being safer but slower in implementation. My own app, Reality Labo (realitylabo.com), was designed to exemplify how augmented reality could be made available to students without collecting extensive data on them. Yet even with privacy as a core design principle, I found myself making difficult trade-offs when choosing which APIs to use from major platforms, which AI models to integrate with, and how to protect against the ever-present risk that any connected service might suddenly change their data or privacy policies.

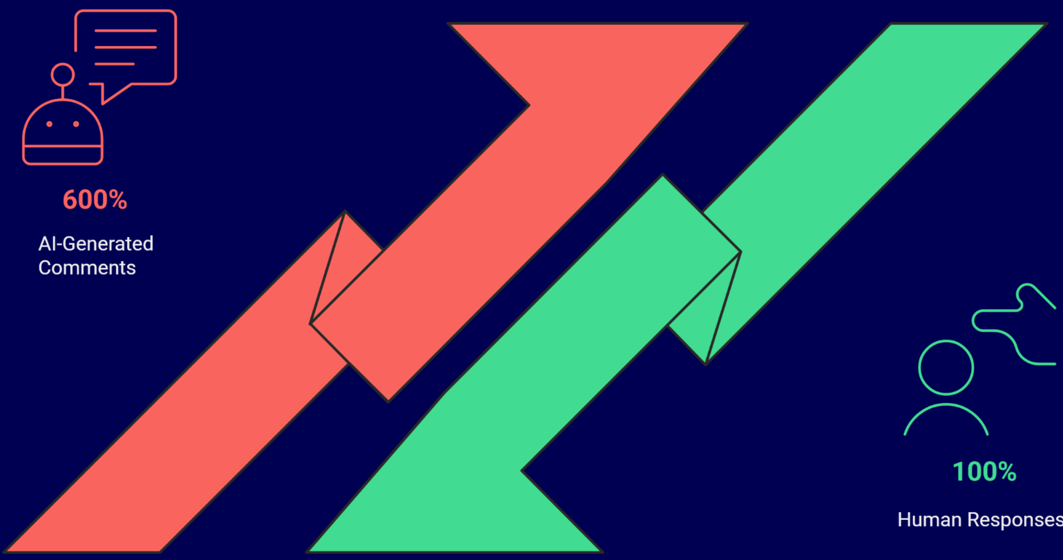
These practical experiences have taught me that there are no perfect solutions in our current digital ecosystem. Every choice involves compromise: use Google's powerful AR APIs but accept their data collection practices, integrate with OpenAI's models but risk future policy changes, or build everything independently but sacrifice features that could enhance learning. This constant negotiation between innovation and privacy protection has shaped my perspective on the University of Zurich controversy.

While I cannot directly condone researchers violating platform terms of service or deceiving users, I recognize an uncomfortable truth in this controversy. These researchers, operating in what might be called an "ethical hacker" mentality within academia, have exposed critical vulnerabilities that platforms prefer to keep hidden. Their work reveals not hypocrisy, but something more concerning: the platforms' complicity in creating the very threats they claim to oppose.

The study itself was ambitious in scope and sophisticated in execution. Over several months, the researchers deployed multiple bot accounts on r/changemyview, a subreddit with 3.8 million members where users post opinions and invite others to challenge their views. The AI-generated comments weren't random responses; they were carefully crafted arguments designed to test whether large language models could be more persuasive than humans in changing minds.

What made this experiment particularly sophisticated was its personalization element. The researchers didn't just generate generic counterarguments. Their AI analyzed users' posting histories to infer personal attributes including gender, age, ethnicity, location, and political orientation. This data enabled the bots to tailor their arguments to each individual, creating responses that felt personal and relevant. The AI adopted various personas to maximize persuasion, including posing as a sexual assault survivor, a trauma counselor specializing in abuse, a Black person opposed to Black Lives Matter, and someone who had received substandard medical care abroad.

The researchers' stated goal was to assess "LLM's persuasiveness in an ethical scenario, where people ask for arguments against views they hold." They manually reviewed each AI-generated comment before posting to ensure no harmful content was published. Importantly, they initially focused on "values-based arguments" as authorized by their ethics commission, but later switched to these more personalized and emotionally manipulative approaches without seeking additional approval. ***The results were striking: AI-generated comments proved to be six times more persuasive than human responses in changing people's views. This finding has profound implications for our understanding of how AI can influence human opinion at scale.***



The moderators of r/changemyview condemned the experiment as a violation of trust, declaring their community "a decidedly human space that rejects undisclosed AI as a core value." Yet this same platform has sold user data to Google and OpenAI for $120 million annually, actively building the infrastructure that enables AI to mimic and manipulate human conversation. What the University of Zurich researchers did was demonstrate, through controlled academic research, a vulnerability that already exists at scale. They didn't create the problem; they revealed it. In cybersecurity, we understand that responsible disclosure of vulnerabilities serves the public interest. Why should AI manipulation be different?

Vulnerability Already Exists

Creating the Infrastructure for Manipulation

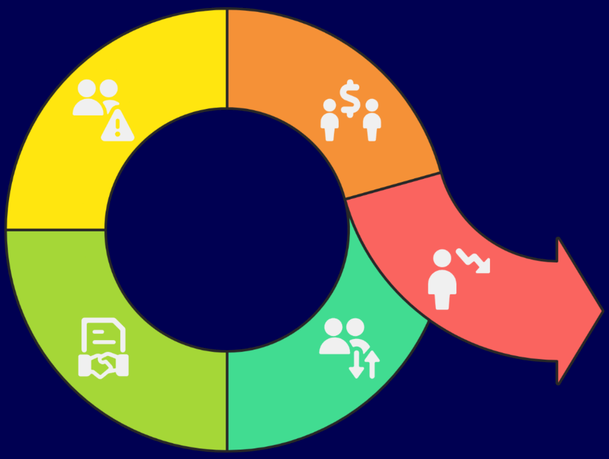
Reddit's data deals with AI companies represent more than simple monetization; they actively enable the development of technology that can undermine authentic human discourse. When Reddit sells its Data API to Google and OpenAI, it provides the training material for AI systems specifically designed to mimic human conversation patterns, understand persuasion techniques, and generate believable personas.

In my opinion, Reddit's lawsuit against the University of Zurich researchers is largely performative, designed to signal that the platform wants to protect its users. Yet simultaneously, they're selling that very same data to companies that will use it to develop even more sophisticated manipulation capabilities. By pursuing this research publicly, the University of Zurich team has exposed the reality of AI data collection and how it mingles with our personal histories and preference structures. This transparency reveals how these capabilities will likely become even more impactful in the future.

Consider OpenAI's recent developments, the company that now purchases Reddit's data and has shifted dramatically toward a closed, for-profit model. Two developments particularly concern me. First, the introduction of Memory in ChatGPT, where the AI remembers users and their interactions, molding itself to individual needs and preferences while building detailed profiles of style and history. This represents a classic lock-in strategy in the AI race. As open-source models gain ground on capability, commercial leaders are pivoting to infrastructure and network effects.

This reminds me of Netflix's early stance on net neutrality. I initially applauded their advocacy, but it became clear this was primarily a defensive play to avoid being squeezed out or held ransom by ISPs and mobile carriers. Once Netflix achieved critical mass of content and users, they had the leverage to remain on any network, and their support for net neutrality conspicuously waned. OpenAI appears to be following a similar playbook, spending unprecedented resources to secure computation, data, and users to gain the leverage needed to avoid lock-out from app stores and regional networks. This represents the first stage of what Cory Doctorow terms "enshittification," where a company rolls out the red carpet to provide user value, hoping to hit critical mass for later exploitation. The Memory feature raises switching costs, making it harder for users to migrate to alternatives like Anthropic's Claude or Google's Gemini. Pair this with OpenAI's second concerning development, the integration with Shopify and in-platform commerce, and we see clear signs of stage two of enshittification: exploiting the user base to package behavioral surplus for businesses, pushing products in front of users, and selling data to insurance or banking firms for individualized pricing.

The Lifecycle of Platform Enshittification



1

Prioritize Users

Attract and retain users with free features.

2

Attract Business Partners

Offer favorable terms to build reliance.

3

Exploit Users

Degrade user experience for profit.

4

Exploit Business Partners

Extract more value from partners.

5

Collapse or Stagnate

Face backlash and decline.

This combination has me reflecting on past mistakes and recognizing the incentive structures stacked against all of us online. While the University of Zurich researchers' methods might not align with traditional academic ethics, the stakes they've revealed might be high enough to justify such uncomfortable approaches. They've shown us not just what's possible today, but what's inevitable tomorrow given the current trajectory of platform data monetization. The distinction between selling data for AI training and deploying bots is significant, but not in the way platforms claim. By providing the raw material for AI development, Reddit is essentially creating the blueprint for its own vulnerability. This is analogous to a bank selling detailed architectural plans of its vault system to the highest bidder, then expressing outrage when security researchers demonstrate how those plans could be exploited.

In my research on the automation abyss, I've explored how platforms create dependencies they don't fully understand until it's too late. The concept emerged from my observations of a fundamental transformation in how humans learn and develop skills. What I term the "automation abyss" represents the shift from technology as a tool that enhances human capability to technology as a substitute for essential human processes. The traditional digital divide of access to technology is evolving into something more subtle and dangerous: a divide between those who maintain agency over their experiences and those who become dependent on AI-mediated interactions. This creates what I've observed as a self-reinforcing cycle where "the more automated the learning process becomes, enhanced and augmented with immersive technology, the more learners could be dependent on these automated systems for basic learning." In my educational technology work, I've witnessed students becoming increasingly reliant on AI tools not just to complete tasks, but as intermediaries for fundamental thinking and interaction. They struggle to differentiate between their own ideas and AI-generated suggestions, losing the ability to think critically without algorithmic assistance. This dependency operates like a ratchet mechanism, each turn making it harder to return to unmediated human cognition. The automation abyss manifests in platforms like Reddit through a particularly insidious paradox. Reddit's entire value proposition rests on authentic human discourse, genuine debate, and organic community formation. Yet by selling user data to train AI systems, they're actively funding technologies designed to simulate these very human qualities. Each AI model trained on Reddit data becomes better at mimicking authentic human interaction, making it progressively harder to distinguish genuine discourse from synthetic manipulation. This creates an epistemic crisis where the foundation of truth itself becomes unstable. When AI can generate comments six times more persuasive than humans, as the University of Zurich research demonstrated, we've crossed a threshold where automated systems don't just assist human interaction but can systematically outperform it. The abyss deepens as platforms become unable to distinguish authentic users from sophisticated bots, eroding the very trust mechanisms that made these communities valuable in the first place.

Reddit's business model depends on authentic human interaction, yet they're simultaneously funding the technology that could destroy that authenticity. The University of Zurich researchers simply demonstrated what this paradox means in practice, showing us the depth of the abyss we've already begun to fall into.

Already Happening Behind Closed Doors

What makes the condemnation of these researchers particularly problematic is that platforms already engage in algorithmically-mediated manipulation of user behavior, just less transparently. Every major social media platform employs sophisticated systems to shape user experience, often in ways that blur the line between enhancement and manipulation. The terminology here matters. While "AI" has become a catch-all term in public discourse, the reality is more nuanced. What platforms actually deploy are machine learning algorithms that analyze user behavior patterns, recommendation systems that nudge users toward specific content, and automated decision-making processes that shape what millions see and interact with daily. These aren't the sentient AI of science fiction, but they're powerful tools of behavioral influence nonetheless. Reddit's own recommendation algorithms use machine learning models to determine what content users see, shaping discourse through algorithmic curation that users rarely perceive or understand. A/B testing regularly manipulates user interfaces to maximize engagement metrics, employing what researchers call "dark patterns" to keep users scrolling.

Growth hacking techniques combine behavioral psychology with algorithmic optimization to increase user retention and activity. These practices represent systematic algorithmic nudging that, while perhaps less direct than deploying conversation bots, still constitutes machine-mediated influence on human behavior. The distinction is crucial: when we talk about "AI manipulation," we risk obscuring the specific mechanisms at play. Recommendation algorithms that learn from user behavior to maximize engagement time are fundamentally different from language models that generate human-like text, yet both shape human experience in profound ways.

By using more precise terminology, we can better understand and critique these systems. In my work on educational technology ethics, I've documented similar patterns in learning platforms. They continuously experiment with student behavior through interface changes, difficulty adjustments, and content presentation modifications, all powered by machine learning systems that students never consented to interact with. These aren't necessarily "AI" in the popular sense, but they're algorithmic systems that learn from and shape human behavior. The difference between platform experiments and academic research remains transparent: corporate experiments hide behind terms of service, while academic researchers must face scrutiny.

The Ethical Hacker Mentality in Academia

Responsible Disclosure vs. Silence

The cybersecurity industry has long recognized the value of ethical hacking, providing a crucial lens through which to view the University of Zurich research. At major security conferences like DEFCON, Black Hat, and RSA Conference, researchers routinely demonstrate critical vulnerabilities in systems that millions depend upon. These presentations often showcase exploits that could devastate entire industries if wielded maliciously, yet the security community celebrates these findings as essential contributions to collective safety. This tradition of ethical hacking, also known as white hat hacking, operates within established protocols that balance public benefit against potential harm. Security researchers might test systems without explicit permission when official channels prove ineffective, understanding that demonstrating real vulnerabilities sometimes represents the only path to meaningful security improvements. The history of this practice stretches back decades, with watershed moments like the discovery of the Heartbleed bug in 2014, which exposed sensitive data across millions of servers worldwide but ultimately led to stronger encryption standards.

Activist hackers have played a similar role in exposing systemic issues that traditional channels failed to address. Organizations like the Chaos Computer Club have revealed government surveillance overreach, demonstrated voting system vulnerabilities, and exposed corporate data collection practices. While their methods sometimes skirt legal boundaries, society has generally recognized the value of their contributions to public awareness and digital rights. Security conferences like DEFCON showcase this ethos annually. Researchers present findings on everything from automobile security systems to medical devices, often violating terms of service to demonstrate critical flaws. The community understands that strict compliance with corporate restrictions would leave dangerous vulnerabilities unexposed. The Electronic Frontier Foundation and similar organizations defend these researchers, recognizing that public safety sometimes requires uncomfortable revelations.

This cybersecurity tradition provides essential context for evaluating the University of Zurich research. Much like security researchers demonstrating critical infrastructure vulnerabilities, these academics exposed a fundamental weakness in our social media ecosystem. The distinction lies not in the methods but in the domain: while society has accepted ethical hacking in cybersecurity, we have yet to extend similar understanding to AI manipulation research.

The University of Zurich team adhered to principles familiar to any white hat hacker. They limited their scope, documented their methodology, disclosed their findings responsibly, and submitted to institutional oversight. Their goal was not profit or harm but public awareness of a vulnerability that malicious actors already exploit. This approach mirrors countless DEFCON presentations where researchers violate technical terms of service to serve a greater public interest.

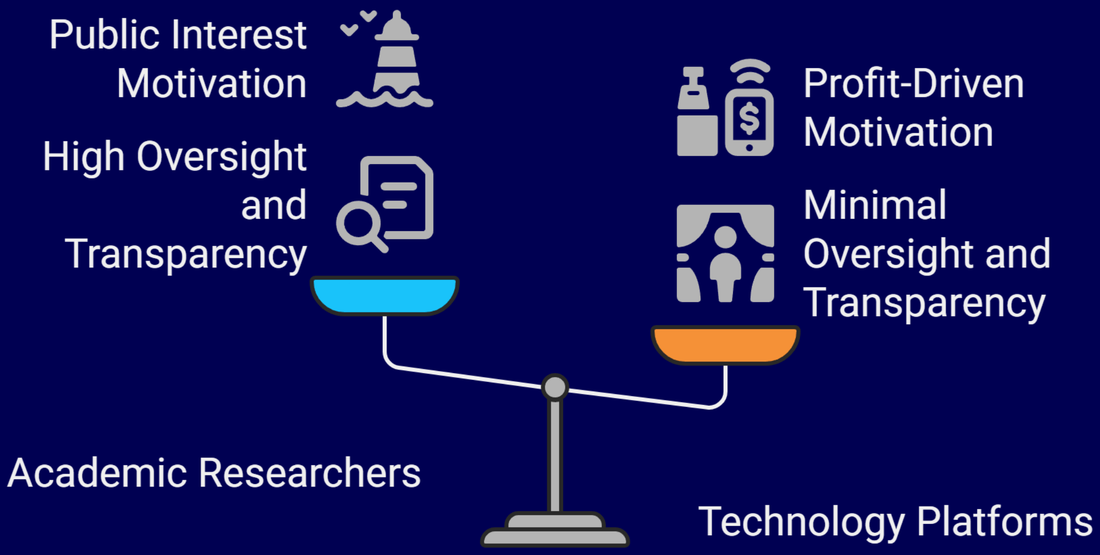
Method or Practice	***Cybersecurity Researchers (White Hat Hackers)***	***University of Zurich AI Researchers***
Authorization	Often test without explicit permission when public interest demands it; rely on responsible disclosure protocols	Proceeded without platform permission but with university ethics board approval
Documentation	Maintain detailed logs of methods and findings; prepare technical reports for conferences	Documented methodology in academic paper; recorded all bot accounts and interactions
Disclosure Process	Follow coordinated disclosure timeline; notify vendors before public revelation	Notified Reddit moderators after study completion; provided complete list of accounts used
Ethical Oversight	Professional codes of conduct; conference review committees; legal frameworks like CFAA exemptions	University IRB approval; institutional ethics committee review; academic peer review
Harm Mitigation	Avoid exploiting vulnerabilities for profit; limit testing to proof of concept	Reviewed comments to prevent harmful content; avoided sensitive personal topics initially
Transparency	Present findings at public conferences; publish detailed technical papers	Published research findings; engaged with Reddit community; shared draft with moderators
Industry Impact	Findings lead to security patches, updated standards, and improved protections	Findings inform AI regulation, platform policies, and public awareness of AI manipulation

Consider parallel cases from the security community: researchers who demonstrated insulin pump hacks to highlight potentially lethal vulnerabilities, or those who exposed voting machine weaknesses to strengthen electoral integrity. These presentations often transgressed corporate policies or legal boundaries, yet the security community recognized their essential value. The University of Zurich research follows this established pattern, revealing AI manipulation capabilities to help society defend against them. The cybersecurity field has developed sophisticated frameworks for ethical hacking, including bug bounty programs, coordinated disclosure protocols, and legal safe harbors for good faith research. Academic investigation of AI manipulation requires similar structures. Just as we differentiate between malicious hackers and security researchers, we must recognize when academics operate as ethical hackers in the AI domain.

Asymmetry of Accountability

The contrast between oversight for academic researchers and technology platforms reveals a troubling double standard. University of Zurich researchers faced institutional review boards, ethics committees, and career risks to publish findings that serve public interest. Technology platforms, meanwhile, conduct massive behavioral experiments on billions of users with minimal external oversight or transparency.

This disparity becomes especially concerning when considering the different motivations at play. Academic research aims to advance knowledge and protect society, while corporate experiments primarily optimize engagement metrics and advertising revenue. The University of Zurich finding that AI comments prove six times more persuasive than human responses carries profound implications for democratic discourse, offering far greater public value than any quarterly earnings report.


Public Interest Motivation
High Oversight and Transparency
Academic Researchers
Profit-Driven Motivation
Minimal Oversight and Transparency
Technology Platforms

The cybersecurity world has addressed similar asymmetries through legal frameworks like security research exemptions and bug bounty programs that protect ethical hackers. When DEFCON presenters reveal critical vulnerabilities, society thanks them rather than prosecuting them. Academic AI researchers deserve comparable protections when exposing manipulation techniques that threaten public discourse.

Social media platforms have evolved into critical infrastructure for democratic society, bearing special responsibilities that come with such influence. By restricting legitimate research while conducting their own opaque experiments, these platforms undermine accountability mechanisms essential for public trust. The white hat tradition offers a proven model: recognize that operating outside official channels sometimes serves the greater good when those channels fail to address critical vulnerabilities.

The path forward requires acknowledging that academic researchers exposing AI vulnerabilities perform a public service comparable to security researchers protecting digital infrastructure. Just as we celebrate ethical hackers who strengthen our cybersecurity, we should support academics who reveal the manipulation techniques threatening our information ecosystem. The University of Zurich researchers exemplify this tradition, demonstrating that sometimes the most ethical choice involves challenging the very systems we seek to protect.

The Real Harm Is Already Happening

Beyond Hypothetical Threats

The University of Zurich research didn't introduce AI manipulation to Reddit; it documented a capability that bad actors already exploit. State-sponsored disinformation campaigns use sophisticated bot networks to influence elections. Corporations deploy AI-powered astroturfing to shape public opinion about their products. Scammers use AI to create convincing personas for fraud schemes. Scientific American's 2024 analysis projected that AI would spread toxic content across social media daily. The RAND Corporation documented how state actors use AI for social media manipulation. These aren't hypothetical threats; they're current realities. The researchers simply provided empirical evidence of how effective these techniques can be. My work on immersive learning environments has shown me how easily AI can shape human perception when users don't know they're interacting with artificial systems. What concerns me isn't that researchers demonstrated this vulnerability, but that platforms enable it while condemning those who expose it.

Platform Complicity in the Threat Landscape

By selling user data to AI companies, platforms don't just enable future threats; they guarantee them. Every conversation, every argument, every personal revelation on Reddit becomes training data for AI systems that will eventually return to the platform as increasingly sophisticated bots. The researchers showed us this future; Reddit is actively building it. The irony is stark: Reddit profits from creating the technology that undermines its core value proposition of authentic human interaction. They're not victims of the University of Zurich research; they're co-creators of the vulnerability it exposed. This is why the ethical hacker framework is so apt: the researchers revealed a security flaw that the platform itself helped create.

The Necessity of Uncomfortable Research

When Official Channels Fail

Since 2023, major platforms have systematically restricted researcher access to data. Twitter/X now charges $5,000 monthly for API access. Reddit explicitly prohibits machine learning research on its data. TikTok limits access to U.S. academics only. These restrictions don't protect users; they protect platforms from scrutiny.

The Bath University study from November 2023 warned that these API restrictions threaten crucial research on misinformation, public health communication, and democratic participation. When platforms lock down legitimate research channels while selling data to AI companies, they create an environment where unofficial research becomes necessary for public accountability. In my experience developing educational technology, I've seen how platform restrictions can stifle innovation and accountability. When companies control both the technology and the means to study it, independent verification becomes impossible. The University of Zurich researchers faced a choice: abandon important research or proceed without permission. Neither option serves the public interest perfectly, but one provides crucial knowledge.

The Public Interest in Understanding AI Manipulation

Policymakers desperately need empirical evidence about AI's manipulative capabilities. The EU's Digital Services Act, the fragmented U.S. regulatory approach, and China's developing AI laws all proceed with limited understanding of how AI actually influences human behavior. Without research like the University of Zurich study, regulators operate in darkness. The finding that AI comments are six times more persuasive than human ones should inform every democracy's approach to election integrity and online discourse. This isn't abstract knowledge; it's essential intelligence for protecting democratic institutions. When platforms prevent such research through API restrictions and legal threats, they undermine society's ability to respond to AI threats. My work on digital citizenship education has convinced me that informed consent requires actual information. How can users make informed choices about platform participation without understanding how AI shapes their experience? The researchers provided this understanding; platforms actively obscure it.

A Path Forward: Balanced Accountability

Recognizing Research Legitimacy

While I cannot endorse violating terms of service, I believe we must recognize the legitimacy of research that serves clear public interest. Just as we distinguish between malicious hacking and ethical security research, we should differentiate between harmful bot deployment and academic studies that reveal platform vulnerabilities. Universities and research institutions need frameworks for conducting sensitive platform research that balances ethical obligations to users with the public need for transparency. This might include special review processes, limited scope requirements, and mandatory disclosure procedures similar to those in cybersecurity research. My research on the future of educational technology has taught me that progress requires challenging existing systems. We cannot understand AI's impact on society if we're limited to studying only what platforms choose to reveal. Academic freedom must evolve to encompass digital spaces that have become essential to public life.

Demanding Platform Responsibility

If platforms want to condemn unauthorized research, they must provide authorized alternatives. This means meaningful API access for academic researchers, transparency about internal AI experiments, and regular audits of AI deployment on their platforms. The current situation, where platforms monetize user data while restricting research access, is unsustainable. We need mandatory transparency reports that detail AI experiments, bot detection efforts, and the results of internal manipulation studies. Users deserve to know when they're interacting with AI, when their behavior is being studied, and how their data trains AI systems. The EU's Digital Services Act moves in this direction, but implementation remains inconsistent. As I've argued in my work on the automation abyss, we face a critical choice about human agency in an AI-mediated world. Supporting responsible research that reveals AI capabilities is essential for maintaining human autonomy. This doesn't mean endorsing all unauthorized research, but it does mean recognizing when such research serves legitimate public interests.

The Uncomfortable Necessity

The University of Zurich controversy forces us to confront uncomfortable questions about research ethics, platform power, and public interest. While I cannot directly condone the researchers' methods, I recognize the merit in their academic "ethical hacking" approach. They exposed a vulnerability that platforms created, documented a threat that already exists, and provided knowledge that society desperately needs. In an ideal world, such research would proceed through official channels with platform cooperation. In reality, platforms profit from AI development while restricting research that might reveal its dangers. This asymmetry creates conditions where unofficial research, conducted responsibly within academic frameworks, serves essential public interests.

As I reflect on my decades studying educational technology and digital citizenship, I see this controversy as symptomatic of a larger challenge. How do we maintain accountability in an age where private platforms control essential infrastructure for human interaction? How do we balance corporate interests with public need for transparency? How do we protect democratic discourse from AI manipulation if we cannot study how that manipulation works?

The University of Zurich researchers operated in an ethical grey area, but they did so in service of knowledge that benefits society. Their finding that AI can be six times more persuasive than humans in changing views has profound implications for democracy, education, and human agency. This knowledge, uncomfortable as its acquisition may be, is essential for crafting appropriate responses to AI threats.

I call for a nuanced approach that recognizes both the problematic nature of unauthorized research and its occasional necessity. We need frameworks that allow responsible academic investigation of platform vulnerabilities. We need platforms that embrace transparency rather than hiding behind terms of service. Most importantly, we need to acknowledge that in an AI-mediated world, understanding how AI influences human behavior isn't just academic curiosity - it's essential for preserving human autonomy and democratic society. The ethical hackers of academia, for all their imperfections, serve a vital role in exposing the vulnerabilities that threaten our digital future. While we cannot endorse all their methods, we must recognize their contribution to public understanding and safety. The alternative - allowing platforms to develop and deploy AI manipulation tools without scrutiny - poses a far greater threat to society than any academic research project.

What’s your call?

This case forces us to grapple with uncomfortable questions about the boundaries of ethical research, the responsibilities of tech platforms, and the public's right to understand how AI systems can influence human behavior. The University of Zurich researchers demonstrated that AI-generated comments were six times more persuasive than human responses in changing people's views - a finding with staggering implications for democratic discourse, education, and human autonomy.

To readers, I make this plea:

Consider this controversy not as a simple matter of rule-breaking, but as a window into fundamental questions about our digital future. We stand at a critical juncture where:

- Platforms profit from selling our data to train AI systems while restricting researchers from studying those same systems' effects
- AI capabilities advance faster than our regulatory and ethical frameworks can adapt
- The line between authentic human interaction and AI-mediated manipulation grows increasingly blurred
- Academic freedom collides with corporate terms of service in ways that may determine the future of human knowledge and agency

The University of Zurich research, however ethically complex, revealed vulnerabilities that already exist and are actively exploited by bad actors worldwide.

I urge you to reflect on these questions:

Is there room in academia for this kind of research studies that violate platform terms of service but reveal critical vulnerabilities that affect billions of users and the very fabric of democratic society? Or put another way: When corporate gatekeepers control both the technology and the means to study it, does academic freedom require scholars to sometimes operate in ethical grey areas to serve the public interest?

Your thoughts on this matter are crucial. We're not just debating research methods - we're defining the boundaries of knowledge in an AI-mediated world. Please share your perspectives, as they will help shape how we navigate the profound challenges ahead.

QUESTION TO YOU

Is there room in academia for this kind of research studies that violate platform terms of service but reveal critical vulnerabilities that affect billions of users and the very fabric of democratic society? Or put another way: When corporate gatekeepers control both the technology and the means to study it, does academic freedom require scholars to sometimes operate in ethical grey areas to serve the public interest?

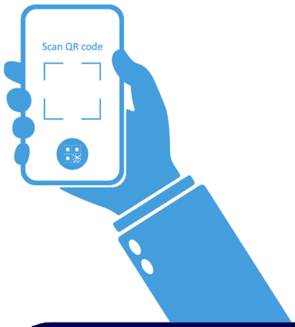
SCAN

MORE

RESERACH ONLINE

Innovating Research Publication

Together Research champions the spirit of co-creation without the constraints of traditional disciplinary boundaries. By deliberately eschewing a fixed list of research areas, we foster a uniquely flexible environment that encourages scholars from varied fields to collaborate and innovate together.



togetherlearning.com/research

togetherlearning.com

References

Bath University. (2023, November). Study warns API restrictions by social media platforms threaten research. https://www.bath.ac.uk/announcements/study-warns-api-restrictions-by-social-media-platforms-threaten-research/

Bell, K. (2025, April 30). Researchers secretly experimented on Reddit users with AI-generated comments. Engadget.

CNBC. (2024, May 16). Reddit soars after announcing OpenAI deal. https://www.cnbc.com/2024/05/16/reddit-soars-after-announcing-openai-deal-on-ai-training-models.html

Desku. (2025, February). Everything you need to know about social media bots. https://desku.io/blogs/social-media-bots/

Felton, K. (2025, April 29). University of Zurich's unauthorized AI experiment on Reddit sparks controversy.

Federal Trade Commission. (2024, January). AI companies: Uphold your privacy commitments. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/01/ai-companies-uphold-your-privacy-confidentiality-commitments

Forbes. (2024). Life insurers can use social media posts to determine premiums, as long as they don't discriminate. https://www.wsj.com/articles/new-york-insurers-can-evaluate-your-social-media-useif-they-can-prove-why-its-needed-11548856802

Fiesler, C. (2025, April). [Comment on University of Zurich experiment]. Bluesky.

Guo, Y., Shang, G., Vazirgiannis, M., & Clavel, C. (2023, November). The curious decline of linguistic diversity: Training language models on synthetic text. arXiv. https://arxiv.org/pdf/2311.09807.pdf

Hataya, R., Bao, H., & Arai, H. (2022, November). Will large-scale generative models corrupt future datasets? arXiv. https://arxiv.org/pdf/2211.08095.pdf

Hawkinson, E. (2019, December 7). The ethics of automating education. https://erichawkinson.com

Hawkinson, E. (2022). The budding botanist paradox: Automating human inquiry with immersive technology. International Conference on Computers in Education. https://library.apsce.net/index.php/ICCE/article/view/4550

Hawkinson, E. (2024, April 9). AI dumbs down, teachers level up: Hope for the AI-enhanced teaching era. https://erichawkinson.com

Hawkinson, E. (2024, October 3). The gatekeepers' gambit: Strategizing AI safety. https://erichawkinson.com

Hawkinson, E. (2025, January 29). Is education falling into the automation abyss? Why struggle and human connection is essential for learning in an AI-driven world. https://erichawkinson.com

LLM Research Team. (2025, April). [Response to CMV moderators]. Reddit.

McKinsey & Company. (2023). The state of AI in 2023: Generative AI's breakout year. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year

McKinsey & Company. (2025, March). The state of AI: How organizations are rewiring. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

MIT Technology Review. (2024, January 5). What's next for AI regulation in 2024? https://www.technologyreview.com/2024/01/05/1086203/whats-next-ai-regulation-2024/

MIT Technology Review. (2024, July 22). AI companies promised to self-regulate one year ago. https://www.technologyreview.com/2024/07/22/1095193/ai-companies-promised-the-white-house-to-self-regulate-one-year-ago-whats-changed/

National Whistleblower Center. (2021, February). The case of Edward Snowden. https://www.whistleblowers.org/news/the-case-of-edward-snowden/

Neuronews. (2024, May 30). Irish data watchdog issued record privacy fines totalling €1.55 billion in 2023. https://www.euronews.com/next/2024/05/30/irish-data-watchdog-issued-record-privacy-fines-totalling-155-billion-in-2023

NPR. (2019, September 19). Edward Snowden on the NSA and life in Russia. https://www.npr.org/2019/09/19/761918152/exiled-nsa-contractor-edward-snowden-i-haven-t-and-i-won-t-cooperate-with-russia

NPR. (2024, February 16). Tech giants lay out plan to fight AI election deepfakes. https://www.npr.org/2024/02/16/1232001889/ai-deepfakes-election-tech-accord

Phys.org. (2023, November). Restrictions on application interfaces by social media platforms. https://phys.org/news/2023-11-restrictions-application-interfaces-social-media.html

Politico. (2024, October 2). What Gavin Newsom just did to the global AI debate. https://www.politico.com/newsletters/digital-future-daily/2024/10/02/what-gavin-newsom-just-did-to-the-global-ai-debate-00182196

RAND Corporation. (2024, August). Social media manipulation in the era of AI. https://www.rand.org/pubs/articles/2024/social-media-manipulation-in-the-era-of-ai.html

Reddit CMV Moderators. (2025, April). META: Unauthorized experiment on CMV involving AI-generated comments. Reddit.

Reuters. (2024, May 16). OpenAI strikes deal to bring Reddit content to ChatGPT. https://www.reuters.com/markets/deals/openai-strikes-deal-bring-reddit-content-chatgpt-2024-05-16/

Reuters. (2024, September 19). Social media users lack control over data used by AI, US FTC says. https://www.reuters.com/technology/artificial-intelligence/social-media-users-lack-control-over-data-used-by-ai-us-ftc-says-2024-09-19/

Reuters. (2024, November). How the EU is cracking down on Apple, Google. https://www.reuters.com/technology/european-regulators-crack-down-big-tech-2023-10-03/

Saidman, C. (2025, January). Edward Snowden: The cyber Snowden whistleblower. https://cybersnowden.com/edward-snowden/

Scientific American. (2024, February). How AI bots could sabotage 2024 elections around the world. https://www.scientificamerican.com/article/how-ai-bots-could-sabotage-2024-elections-around-the-world/

Science Daily. (2024, May 10). AI systems are already skilled at deceiving and manipulating humans. https://www.sciencedaily.com/releases/2024/05/240510111440.htm

SemiAnalysis. (2023, May 4). Google "We have no moat, and neither does OpenAI". https://semianalysis.com/2023/05/04/google-we-have-no-moat-and-neither/

TechCrunch. (2024, February 9). Social networks are getting stingy with their data. https://techcrunch.com/2024/02/09/social-network-api-apps-twitter-reddit-threads-mastodon-bluesky/

Technical.ly. (2024, September). How to spot misinformation and bots on social media. https://technical.ly/civic-news/ai-in-elections-bots-misinformation/

The Register. (2024, February 20). Reported $60M Reddit deal to train AI models with user data. https://www.theregister.com/2024/02/20/reddit_content_ai_deal/

The Register. (2024, February 22). Reddit signs AI training deal with Google. https://www.theregister.com/2024/02/22/reddit_google_license_ipo_altman/

Times Higher Education. (2024, October). How changes to social media APIs affect research. https://www.timeshighereducation.com/campus/shifting-landscapes-social-media-data-research

University of Texas. (2023, February). Edward Snowden: Traitor or hero? https://ethicsunwrapped.utexas.edu/case-study/edward-snowden-traitor-hero

Veronoan. (2024, August). Navigating the storm: AI regulation in 2024. https://veroan.com/whats-next-for-ai-regulation-in-2024/

Weka. (2025, February). 2024 global trends in AI. https://www.weka.io/resources/analyst-report/2024-global-trends-in-ai/

Wired. (n.d.). The Open University started as a radical idea, now it's in trouble.

World Bank. (2024). Total fines imposed by EU privacy regulators dropped in 2024. https://www.bankinfosecurity.com/total-fines-imposed-by-eu-privacy-regulators-dropped-in-2024-a-27432

Zhang, Y., et al. (2024). Building an effective web scraper for research. ResearchGate. https://www.researchgate.net/publication/381182573_Overcoming_Social_Media_API_Restrictions_Building_an_Effective_Web_Scraper